Final Report for the Project "Robust Rapid Change-Point Detection in Multi-Sensor Data Fusion and Behavior Research" (FA9550-08-1-0376)

Yajun Mei (PI)

February 25, 2011

Contents

1 Introduction

The overall goal of our AFOSR sponsored research program is to develop a general and systematic foundation and methodologies for robust rapid change-point detection in the context of fusing noisy data from heterogeneous networked sensors, and apply them to model behavior data in experiments with uncertain onset time of stimulus. Our results offer a deeper understanding how humans process information in dynamic environments under time pressure, and provide new mathematical tools to model behavior data.

Rapid change-point detection, or sequential change-point detection, has a variety of applications such as industrial quality control, signal detection and biosurveillance. The classical version of this problem, where one monitors a single *univariate* independent and identically distributed (i.i.d.) data stream, is a well-developed area, and many classical schemes have been developed such as the Shewhart's chart, moving average control charts, Page's CUSUM

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **28 FEB 2011**	2. REPORT TYPE	3. DATES COVERED **01-07-2008 to 30-11-2011**

4. TITLE AND SUBTITLE
Robust Rapid Change-Point Detection In Multi-Sensor Data Fusion And Behavior Research

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Georgia Institute of Technology,765 Ferst Drive NW,Atlanta,GA,300332

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The overall goal of our AFOSR project is to develop a general and systematic foundation and methodologies for robust rapid change-point detection in the context of fusing noisy data from heterogeneous networked sensors, and apply them to model behavior data in experiments with uncertain onset time of stimulus. Specifically, we derive quickest detection schemes that are asymptotically optimal under different scenarios and spurred by the two-choice experiments in which the subject does not know the time of occurrence of the signal and is allowed to make decisions before the signal appears, we extend Ratcliff's diffusion model to the 2-CUSUM process model by emphasizing natural connections to the sequential change-point detection problems in statistics

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**89**	

procedure, Shiryaev-Roberts procedure and scan statistics. In nowadays information technology infrastructure age, rapid change-point detection finds challenging new applications due to its ubiquitous nature. However, we need to extend the theories and methodologies beyond the classical i.i.d. models.

Our AFOSR sponsored research program focuses on two specific applications. The first is the multi-sensor data fusion problem, which has many important applications, including cognition, psychology, sensor network systems, surveillance systems, and economic theory of teams. Practically, a multi-sensor system can be a collective of intelligent sensors, robots or agents acting together to solve a common task, and its motivation is to mimic the ongoing cognitive process used by humans that integrates data continually from their senses to make inferences about the external world. The second involves in psychology and neuroscience, where researchers model the human (and animal) decision making process as a form of sequential sampling process. We illustrate that rapid change-point detection methodologies provide new mathematical tools for modelling.

This report will furnish the successes and achievements accomplished under AFOSR funding.

2 Overview of Accomplishments

2.1 Efficient Scalable Global Detection Schemes

In the modern information age, one often faces the need to monitor a large number of noisy data streams with the aim of offering the potential for early detection of a “trigger” event. In the literature, two special categories of global monitoring have gained a lot of attention: one is consensus detection in which the occurring event is assumed to affect all data streams abruptly and simultaneously, and the other is the case when one and only one data stream is known to be affected by the occurring event, e.g., multi-channel detection in signal detection.

In our research, we consider a general scenario when the occurring event may affect different data streams differently. For instance, we may know that the event can affect 10 out of 100 data streams but we do not know which 10 data streams will be affected. In addition, the event could have an immediate or delayed impact on affected data streams. One example is biosurveillance when a disease outbreak may first occur in some unknown local cities or counties and then spread out to other regions. Another example occurs in target detection when one uses several different sensor types such as radar, sonar, infrared and magnetic to detect a target, and it is possible that some (but not all) sensors can provide earlier information about the target.

Ideally, one would like to develop a single global monitoring scheme that can detect the occurring event as quickly as possible while controlling the system-wise global false alarm rate. While such global monitoring schemes can be found by the standard statistical methods such as generalized likelihood ratios or mixture likelihood ratios from the theoretical viewpoint, they are generally infeasible from the computational viewpoint since they would require us to do an exhaustive search over all possible post-change hypotheses or over all possible combinations of affected data streams. Hence, it is desirable to find a global monitoring scheme that is not only efficient, but also *scalable* in the sense of being able to be implemented to monitor a large number of data streams over a long period of time.

To develop a scalable global scheme, one intuitively appealing approach is to monitor each local data stream locally through some classical computationally efficient schemes and then combine all local schemes together to produce a single global scheme. Unfortunately, little research has been done on how to combine all local schemes together to ensure efficiency, and to the best of our knowledge, only a naive approach has been proposed so far. The naive approach is to raise an alarm at the global level whenever any local scheme raises a local alarm. This naive approach is very effective if one or very few data streams are affected, but it may lose efficiency when several or more data streams can provide information to the occurring event. For the purpose of comparison, below the global scheme corresponding to this naive approach will be referred as a "MAX" scheme, since it can be thought of raising a global alarm if the *maximum* of local monitoring statistics is too large, where, if necessary, one can normalize the local schemes so that the local detection thresholds are the same.

Our research offers new approaches to combine all local schemes together to produce an efficient scalable global scheme for rapid change-point detection. In Mei (2010), we propose to raise an alarm at the global level if the *sum* of local monitoring statistics (e.g., local CUSUM statistics in the logarithm scale) is too large, and theoretical analysis and numerical simulations show that the corresponding "SUM" scheme is efficient when the number of affected data streams is very large.

In Mei (2011), we take a further step to monitor a large number of data streams via *thresholding*. The fundamental idea is to raise a global alarm based on the sum of those local detection statistics (e.g., local CUSUM statistics) that are "large" under either top-r thresholding rules or hard thresholding or both. It turns out that the corresponding global schemes include both MAX and SUM schemes as special cases and possess certain asymptotic optimality properties under proper criteria. It is worth pointing out that the thresholding idea has been applied for high-dimensional data in the off-line statistical inference literature to improve power or efficiency, but our application to rapid change-point detection is new.

Mathematically, assume that in a system, one is monitoring K data streams over time n, say, $\{X_{k,n}\}_{n=1}^{\infty}$ for $k = 1, \ldots, K$. Initially, the system is "in control" and the distribution of the $X_{k,n}$'s is f_k for the k-th data stream. At some *unknown* time ν, a "trigger" event occurs and affects the system in the sense that the density function of the observations $X_{k,n}$'s changes from f_k to another given density g_k at time $\nu_k \geq \nu$. Without loss of generality, we assume that $\min_{1 \leq k \leq K} \nu_k = \nu$, since otherwise the system is affected by the event only after the time $\nu' = \min_{1 \leq k \leq K} \nu_k$, which can be treated as the new change-point. It is desired to utilize the observed data streams $X_{k,n}$'s to raise an alarm at the global level as soon as the event occurs so that one can take appropriate action.

A standard minimax formulation is to minimize the detection delay subject to a false alarm constraint. The latter is typical of the form

$$\mathbf{E}^{(\infty)}(T) \geq \gamma, \tag{1}$$

where T is the stopping rule associated with the detection scheme, $\gamma > 0$ is a pre-specified constant, and $\mathbf{E}^{(\infty)}$ denote the expectation when there are no changes. In the literature $\mathbf{E}^{(\infty)}(T)$ is often called as the average run length to false alarm. The definition of detection delay is a little more complicated, as it needs to take into account the uncertainty of the change-point ν. One widely used definition is the following "worst case" detection delay

$$\mathbf{D}(T) = \sup_{1 \le \nu < \infty} \operatorname{ess\,sup} \mathbf{E}^{(\nu)} \Big((T - \nu + 1)^{+} \Big| \mathcal{F}_{\nu-1} \Big), \tag{2}$$

where $x^{+} = \max(x, 0)$, $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$ denotes past global information at time ν, and $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is past local information for the k-th data stream.

Before discussing the existing schemes for globally monitoring, let us consider the local schemes for monitoring a single data stream, say, the kth data stream. Such a problem has been well studied in the sequential change-point detection literature, and one efficient local scheme is Page's CUSUM procedure which raise a local alarm as soon as the local CUSUM statistic exceeds some pre-specified threshold, where the local CUSUM statistic for the kth data stream at time n is given by

$$W_{k,n} = \max \Big\{ 0, \ \max_{1 \le \nu \le n} \sum_{i=\nu}^{n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \Big\}.$$

In practice, the local CUSUM statistic $W_{k,n}$ can be computed conveniently online via a recursive formula:

$$W_{k,n} = \max \Big(W_{k,n-1} + \log \frac{g_k(X_{k,n})}{f_k(X_{k,n})}, \ 0 \Big), \tag{3}$$

and $W_{k,0} = 0$.

Now let us go back our global monitoring problem. As mentioned before, the naive approach is the "MAX" scheme that raises an alarm at the global level as soon as one of local CUSUM procedures raises a local alarm. Specifically, the "MAX" scheme raises a global alarm at time

$$T_{\max}(c) = \text{ first time } n \ge 1 \text{ such that } \max_{1 \le k \le K} W_{k,n} \ge c, \tag{4}$$

($= \infty$ if such n does not exist) where $c > 0$ is a pre-specified constant chosen to satisfy the false alarm constraint (1).

In Mei (2010), we propose the "SUM" scheme that raises an alarm when the sum of local CUSUM statistics is too large. Specifically, at time n, each data streams calculates its local CUSUM statistic $W_{k,n}$'s as in (3), and then one will raise an alarm at the global level at time

$$T_{\text{sum}}(d) = \text{ first time } n \ge 1 \text{ such that } \sum_{k=1}^{K} W_{k,n} \ge d, \tag{5}$$

($= \infty$ if such n does not exist) where the constant $d > 0$ is some suitably chosen constant. The intuition behind the "SUM" rule is that a large value of local CUSUM statistic indicates a possible local change, and thus the sum of local CUSUM statistics is a reasonable detection statistic if the number of affected data streams is large or completely unknown. Intuitively, the "MAX" scheme $T_{\max}(c)$ in (4) works better when one or very few data streams are affected, whereas the "SUM" scheme $T_{\text{sum}}(d)$ in (5) works better when many data streams are affected, and numerical simulations in Mei (2010) indeed verified this intuition.

Recently, in Mei (2011), motivated by the scenario when one has a prior knowledge that at most r data streams will be affected and by the applications in censoring sensor networks, we propose the following new global schemes:

- **The top-r thresholding scheme.** At each time n, we order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ from largest to smallest: $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$. Then the top-$r$ thresholding scheme raises a global alarm at time

$$N_{top,r}(a) = \text{ first time } n \geq 1 \text{ such that } \sum_{k=1}^{r} W_{(k),n} \geq a. \tag{6}$$

- **The hard thresholding scheme.** At time n, each local data stream calculates its local CUSUM statistic $W_{k,n}$ recursively as in (3). Then at time n, the decision-maker will take a "keep or kill" policy to make a decision based on

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases},$$

where $b_k \geq 0$ is the local censoring (hard threshold) parameter at the k-th data stream. To be more concrete, our proposed hard thresholding scheme raises a global alarm at time

$$N_{hard}(a, b) = \text{ first } n \geq 1 \text{ such that } \sum_{k=1}^{K} W_{k,n} I\{W_{k,n} \geq b_k\} \geq a. \tag{7}$$

A "good" choice of the b_k's is $b_k = \rho_k b$ for $k = 1, \ldots, K$ for some constant $b \geq 0$ where

$$\rho_k = \frac{I(g_k, f_k)}{\sum_{k=1}^{K} I(g_k, f_k)}$$

can be thought of as the weight of the k-th data stream in the overall final decision, and $I(g_k, f_k)$ is the Kullback-Leibler information number defined by

$$I(g_k, f_k) = \int \log \frac{g_k(x)}{f_k(x)} g_k(x) d\mu(x).$$

- **The Combined Thresholding Scheme.** At time n, each local data stream uses the hard threshold method to generate $U_{k,n} = W_{k,n} I\{W_{k,n} \geq \rho_k b\}$, and the decision maker uses the top-r thresholding rule to combine $U_{k,n}$'s to make a decision. Mathematically, the combined thresholding scheme raises an alarm at the global level at time

$$N_{comb,r}(a,b) = \text{ first } n \geq 1 \text{ such that } \sum_{k=1}^{r} U_{(k),n} \geq a. \tag{8}$$

It is worth mentioning that our research has led to some interesting spin-off results in renewal theory and applied probability. Asymptotic analysis of the properties of our proposed new schemes involves the development of some novel methods to investigate the limiting behavior of $W_n^* = \max_{0 \leq i \leq n} \sum_{k=1}^{K} W_{k,i}$ under *different* scenarios, where $W_{k,i} = \max(W_{k,i-1} + \xi_{k,i}, 0)$ with $W_{k,0} = 0$, and the $\xi_{k,i}$'s are K independent (not necessarily identical) sequences of random variables. The CUSUM-like statistic $W_{k,i}$ and its extreme value process W_n^* play an important role in queue theory, where $W_{k,i}$ is the waiting time of the i-th customer. In the literature the properties of W_n^* with $K = 1$ has been investigated, but that of W_n^* with $K \geq 2$ has not been studied so far. During the process, we find an elegant result on the first ladder epoch τ where all K independent one-dimensional random walks reach recording-setting values simultaneously. Specifically, in Mei (2010), we show that under a general condition, $\mathbf{E}(\tau) = \prod_{k=1}^{K} \mathbf{E}(\tau_k)$, where τ_k is the first weak ladder epoch of the k-th one-dimensional random walk. This simple and interesting result is a new contribution to the well-developed renewal theory.

2.2 Non-homogeneous Poisson

Another topic of our research is concerned with the scenario when the observations are non-homogeneous under the baseline. Motivated by applications in bio and syndromic surveillance, we investigate the problem of detecting a change in the mean of Poisson distributions after taking into account the effects of population size. The specific data motivating our research concerns male thyroid cancer cases (with malignant behavior) in New Mexico during 1973-2005. The data set has been studied before in the biosurveillance literature and is available from the Surveillance, Epidemiology, and End Results (SEER) Program at the National Cancer Institute that collects information on cancer incidence, mortality, and survival from the population-based cancer registries in the United States. Figure 1 plots three different curves related to this data set: (1) yearly total number of cancers with malignant behavior; (2) yearly population size (of males) in New Mexico; and (3) yearly (crude) incidence rate per 100,000 (male) population.

An interesting goal in epidemiology and (bio)surveillance is to determine whether or not the *risk* for male thyroid cancer increases over time. The term *risk* here essentially means the probability of developing thyroid cancer in a given year, which can be characterized by the incidence rate per 100,000 (male) population; see the plot in the bottom panel of Figure 1. Since the binomial distribution can be approximated by Poisson distribution with the

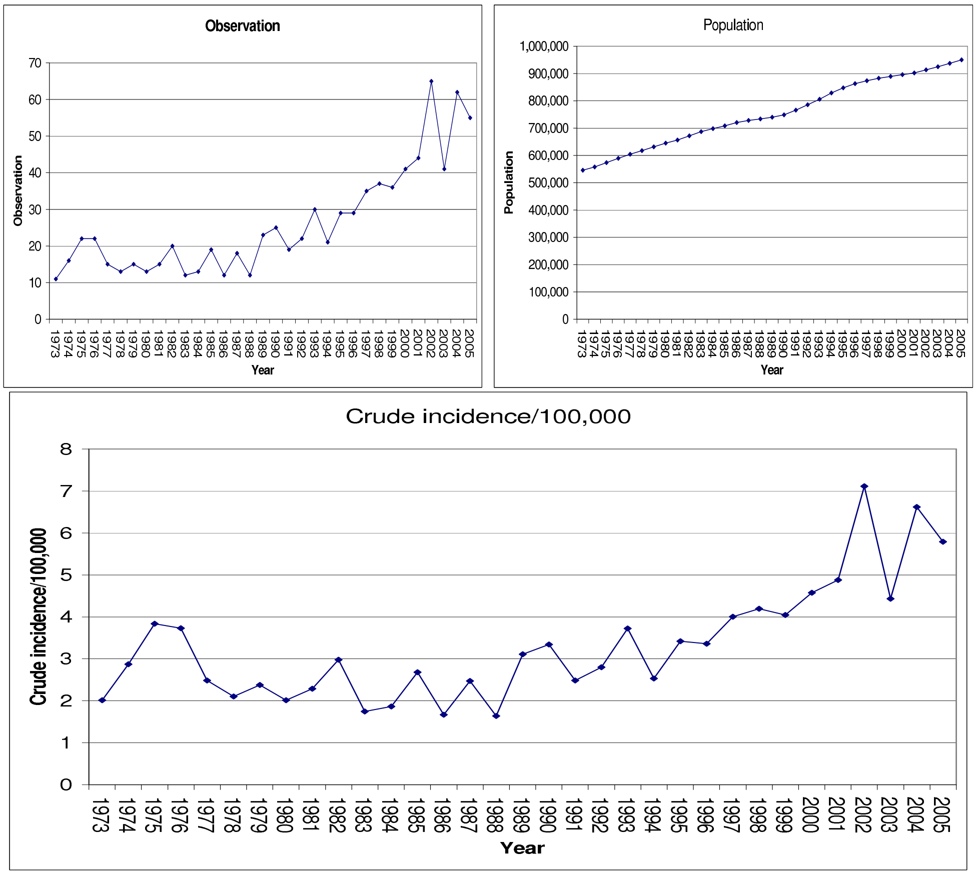

Figure 1: Three time series data of male thyroid cancer in New Mexico during 1973-2005. Top: the left panel plots the total number of male thyroid cancers over years, and the right panel illustrates the trend of the male population. Bottom: the plot is of the crude cancer incidence per 100,000 population over years.

same, we can simply assume that we observe two-dimensional random vectors (l_n, Y_n) over time n, where Y_n has a Poisson distribution with mean $\mu_n = l_n \lambda_n$. Here l_n, Y_n and λ_n can be thought of as the population size (in the units of 100,000 population), the number of disease cases, and the (unobservable true) incidence rate per 100,000 (male) population at the n-th year, respectively.

Under a very simplified setting, the λ_n's, e.g., the incidence rates per 100,000 (male) population, are assumed to change from one value λ_0 to another value λ_1 at some unknown time ν, and we want to detect such a change as soon as possible if it occurs. Note that we are only interested in detecting a change in the risk λ_n's, and the population sizes l_n's can be either pre-specified constants or observable (possibly dependent) random variables whose distributions are nuisance parameters that are left unspecified.

To construct efficient schemes, it is natural to consider the family of generalized likelihood ratio (GLR) schemes. Note that the problems can be thought of as testing the null hypothesis $H_0 : \nu = \infty$ (no change) against the composite alternative hypothesis

$H_1 : 1 \leq \nu < \infty$ (a change occurs), the logarithm of the corresponding GLR statistic of the first n observations, $\{(l_i, Y_i)\}_{i=1}^{n}$, is given by

$$W_n = \max_{1 \leq \nu < \infty} \log \frac{d\mathbf{P}_\nu}{d\mathbf{P}_\infty}\Big((l_1, Y_1), \cdots, (l_n, Y_n)\Big).$$

Now given the l_i's, the Y_i's are conditionally independent with a conditional probability density function (pdf) $f_0(Y_i|l_i) = e^{-l_i \lambda_0}(l_i \lambda_0)^{Y_i}/(Y_i!)$ if $i < \nu$, but with a conditional pdf $f_1(Y_i|l_i) = e^{-l_i \lambda_1}(l_i \lambda_1)^{Y_i}/(Y_i!)$ if $i \geq \nu$. Moreover, the distribution of the l_n's is assumed to be the same under $\mathbf{P}_\infty$ or $\mathbf{P}_\nu$, and for the first n observations, $\{(l_i, Y_i)\}_{i=1}^{n}$, their $\mathbf{P}_\nu$-distribution is the same as their $\mathbf{P}_\infty$-distribution when $\nu > n$, due to the uniqueness of the pre-change distribution. Hence, the logarithm of the GLR statistic can be rewritten as

$$\begin{aligned} W_n &= \max_{1 \leq k \leq n+1} \sum_{i=k}^{n} \log \frac{f_1(Y_i|l_i)}{f_0(Y_i|l_i)} \\ &= \max_{1 \leq k \leq n+1} \sum_{i=k}^{n} \Big[Y_i \log \frac{\lambda_1}{\lambda_0} - l_i(\lambda_1 - \lambda_0) \Big], \end{aligned} \tag{9}$$

where $\sum_{i=n+1}^{n} = 0$. Thus, under our setting, the GLR scheme raises an alarm at time

$$T_{GLR}(a) = \text{ first } n \geq 1 \text{ such that } W_n \geq a, \tag{10}$$

($= \infty$ if such n does not exist), where the constant a is chosen to satisfy the false alarm constraint. For the purpose of online implementation, it is easy to see that W_n in (9) enjoys a recursive formula of the classical CUSUM statistics:

$$W_n = \max\Big\{0, W_{n-1} + \Big[Y_n \log \frac{\lambda_1}{\lambda_0} - l_n(\lambda_1 - \lambda_0)\Big]\Big\}.$$

It is also not difficult to establish the asymptotic optimality properties of the GLR scheme under the classical asymptotic setting.

So far we have "solved" the problem by our favorite GLR methods, but perhaps not solved it in practice. To illustrate that despite its nice asymptotic optimality properties, the GLR scheme may not necessarily be as effective as one expects in application, we propose two ad-hoc methods for comparison.

Intuitively, two features of the GLR scheme seem questionable in the context of non-stationary population sizes: (i) the GLR statistic W_n in (9) assigns the same weight to the individual log-likelihood ratio statistic $\log \frac{f_1(Y_i|l_i)}{f_0(Y_i|l_i)}$ regardless of population size l_i's, although the Y_i's with larger population sizes l_i's surely provide more information, and (ii) the GLR scheme $T_{GLR}(a)$ uses the constant threshold value a over time. Accordingly, we propose two alternative detection schemes to take into account the effects of population sizes.

The first scheme is based on the quasi-log-likelihood ratio statistics that normalize each term $\log \frac{f_1(Y_i|l_i)}{f_0(Y_i|l_i)}$ in (9) by their (conditional) variances, or equivalently, by the population sizes l_i's (up to a constant). This leads to the detection statistic

$$\begin{aligned} \hat{W}_n &= \max_{1 \le k \le n+1} \sum_{i=k}^{n} \frac{1}{l_i} \log \frac{f_1(Y_i|l_i)}{f_0(Y_i|l_i)} \\ &= \max_{1 \le k \le n+1} \sum_{i=k}^{n} \Big[\frac{Y_i}{l_i} \log \frac{\lambda_1}{\lambda_0} - (\lambda_1 - \lambda_0) \Big]. \end{aligned} \tag{11}$$

Thus, for any given constant b, we can define the weighted likelihood ratio (WLR) scheme

$$T_{WLR}(b) = \text{ first } n \ge 1 \text{ such that } \hat{W}_n \ge b. \tag{12}$$

Another motivation for the WLR scheme $T_{WLR}(b)$ in (12) is based on Y_n/l_n, a natural estimator of the *risk* or the disease rate per 100,000 population. To see this, if we pretend that Y_n/l_n is Poisson distributed with mean λ_n (this is not true under our setting, but we can still use it to construct detection schemes), then the problem becomes the classical problem of detecting a change in the Poisson mean from λ_0 to λ_1, and the corresponding GLR (or CUSUM) procedure is just the WLR scheme $T_{WLR}(b)$ in (12).

The second scheme we propose is to use the GLR-based statistic W_n in (9), but with adaptive thresholds to take into account population size effects. Ideally, one would like to use the optimal thresholds or boundaries, say, by some Bayesian or non-Bayesian arguments, but such boundaries seem to be too complicated to derive explicitly. For simplicity, we use the linear boundaries: $l_n c$. Specifically, the proposed adaptive threshold method (ATM) raises an alarm at time

$$T_{ATM}(c) = \text{ first } n \ge 1 \text{ such that } W_n \ge l_n c, \tag{13}$$

for some constant $c > 0$, where W_n is the GLR statistic defined in (9).

It is important to point out that when the population sizes l_n's are equal to a constant $l > 0$, then the three detection schemes, $T_{GLR}(a), T_{WLR}(b)$ and $T_{ATM}(c)$, not only are equivalent (when $a = lb = lc$), but also hold the exact optimality properties of Page's CUSUM procedures for the i.i.d. models. However, when the population sizes l_n's vary, these three schemes are no longer equivalent.

Indeed, numerical simulation studies illustrate that the GLR schemes are at times not as efficient as two families of ad-hoc schemes based on either the weighted likelihood ratios or the adaptive threshold method that adjust the effects of population sizes. Specifically, the GLR scheme is the best scheme with the smallest worst-case detection delay if the population sizes are decreasing, but it is the worst scheme if the population sizes are increasing; the WLR scheme is the worse scheme if the population sizes are decreasing, but the best if the population sizes are increasing. The adaptive threshold scheme $T_{ATM}(c)$ seems to be *robust* in the sense of small detection delays $\overline{\mathbf{E}}_1(T)$, no matter whether the population sizes are increasing or decreasing.

In other words, our simulations suggest that when one has prior information that population sizes are increasing or decreasing, one use the best among these three schemes. When there is uncertainty about the trends of population sizes, one may want to use the adaptive threshold scheme $T_{ATM}(c)$ to take advantage of its robustness properties. In particular, despite its asymptotic optimality properties under the classical asymptotic setting, the GLR scheme indeed can perform very poorly in finite-sample numerical simulations, especially in the typical scenarios of biosurveillance when the population sizes are increasing.

To explain this, in Mei, Han and Tsui (2011), we develop a new asymptotic optimality analysis under a new asymptotic setting that is more suitable to our finite-sample numerical simulations. In addition, we extend our approaches to a general setting with arbitrary probability distributions, as well as to the continuous-time setting involving the multiplicative intensity models for Poisson processes, although further research is needed.

2.3 Decentralized Sequential Multi-Hypothesis Testing

In Wang and Mei (2010), we investigate decentralized sequential multi-hypothesis testing problems. In recent years, the decentralized version of sequential hypothesis testing problems has gained a great amount of attention and has been applied into a wide range of applications such as military surveillance, target tracking and classification, and data filtering. Under a widely used decentralized setting, raw data are observed at a set of geographically deployed sensors, whereas the final decision is made at a central location, often called the fusion center. The key feature here is that raw observations at the local sensors are generally not directly accessible by the fusion center, and the local sensors need to send quantized summary messages (generally belonging to a finite alphabet set) to the fusion center. This is due to limited communication bandwidth and requirements of high communication robustness.

Unfortunately, decentralized sequential hypothesis testing problems are very challenging, and to the best of our knowledge, existing research is restricted to testing two simple hypotheses, and it has been an open problem to find any sort of asymptotically optimal solutions for the decentralized sequential testing problem when testing $M \geq 3$ hypotheses. This is not surprising, because even in the centralized version, it requires sophisticated mathematical and statistical techniques and only asymptotic optimality results are available.

The goal of our research along this direction is to tackle this open problem by developing a class of asymptotically optimal decentralized sequential procedures for testing $M \geq 3$ hypotheses. To do so, a major challenge we need to overcome is finding the "optimal quantizers" that can best send quantized summary sensor messages from the local sensors to the fusion center so as to lose as little information as possible. Intuitively, such a quantizer should depend on the true distribution of the raw data, which is unknown, and thus stationary quantizers are generally not optimal. In addition, since a quantizer can be any measurable function as long as its range is in the given finite alphabet set, it resides in an infinite dimensional functional space. Hence it is essential to investigate the form of the "optimal quantizers" so that one can reduce the infinite dimensional functional space to a finite-dimensional parameter space for the purpose of theoretical analysis and numerical computation. Note that when testing $M = 2$ hypotheses, Tsitsiklis (*IEEE Trans. Commun.*, 1993) and Veeravalli, Basar, and Poor (*IEEE Trans. Inf. Theory*, 1993) showed that the optimal quantizers can be found from the family of monotone likelihood ratio quantizers (MLRQ), whose form is defined up to a finite number of parameters. Unfortunately, such a result does not apply to the case of testing $M \geq 3$ hypotheses. To find the form of the optimal quantizers for multi-hypotheses, we combine three existing methodologies together:

- *two-stage tests* in Stein (*Ann. Math. Statist.*, 1945) and Kiefer and Sacks (*Ann. Math. Statist.*, 1963), or equivalently, tandem quantizers in Mei (*IEEE Trans. Inf. Theory*, 2008);
- *unambiguous likelihood quantizers* (ULQ) in Tsitsiklis (*IEEE Trans. Commun.*, 1993); and
- *randomized quantizers*, see Chernoff (*Ann. Math. Statist.*, 1959) for a closely related topic on randomized experiments.

Mathematically, we assume that a sensor network consists of K local sensors labeled by $S^1, \ldots, S^K$ and a fusion center which makes a final decision when stopping taking observations. At each time step $n = 1, 2, \ldots$, each local sensor S^k observes raw data $\{X_n^k\}$ and sends quantized summary messages $\{U_n^k\}$ to the fusion center. Here the quantized messages $\{U_n^k\}$ are required to belong to a finite alphabet, say, $\{0, 1, \ldots, l^k - 1\}$, due to limited communication bandwidth or requirements of high communication robustness. In other words, the fusion center does not have direct access to the raw data $\{X_n^k\}$, and have to utilize the quantized sensor messages $\{U_n^k\}$ to make a final decision. If necessary, the fusion center can send feedback $\{V_n^k\}$ to the local sensors so as to improve the system efficiency. To be more concrete, we further assume that at time n, for each $k = 1, 2, \ldots, K$, the quantized sensor message at the kth local sensor is of the form

$$U_n^k = \phi_n^k(X_n^k; V_{n-1}^k) \ \in \{0, 1, \ldots, l^k - 1\} \tag{14}$$

where the feedback V_{n-1}^k is defined by

$$V_{n-1}^k = \psi_n^k(U_{[1,n-1]}^1, \ldots, U_{[1,n-1]}^K) \tag{15}$$

and $U_{[1,n-1]}^k = (U_1^k, \ldots, U_{n-1}^k)$ denotes all past local sensor messages. That is, the quantizer ϕ_n^k is a function used by sensor S^k to map the local raw data X_n^k into $\{0, 1, \ldots, l^k - 1\}$, and the choice of ϕ_n^k can depend on the feedback V_{n-1}^k and can be a randomized function (to be discussed later).

In decentralized sequential multihypothesis testing problems, there are M hypotheses regarding the distribution $\mathbf{P}$ of the raw data $\{X_n^k\}$:

$$\mathbf{H}_m : \quad \mathbf{P} = \mathbf{P}_m, \quad m = 0, 1, \ldots, M - 1. \tag{16}$$

Under each $\mathbf{P}_m$, the raw data X_n^k at local sensor S^k are i.i.d. with density $f_m^k(\cdot)$ with respect to a common underlying measure, and the raw data $\{X_n^k\}$ are assumed to be independent among different sensors. Hence the distributions of the raw data under $\mathbf{P}_m$ are completely determined by the K densities: $f_m^1, \ldots, f_m^K$. Below we simply state that the true state of nature is m or $\mathbf{P}_m$ if the hypothesis $\mathbf{H}_m$ is true.

A decentralized sequential test δ consists of a rule to determine the sensor messages, a stopping time N used by the fusion center and a final decision rule $D \in \{0, 1, \ldots, M-1\}$ that chooses one of the M probability measures $\mathbf{P}_m$'s based on the information up to time N at the fusion center. In a Bayesian framework, we assign prior probabilities $\pi = (\pi_0, \ldots, \pi_{M-1})$ to the M hypotheses $\mathbf{H}_0, \cdots, \mathbf{H}_{M-1}$. Let $c > 0$ be the cost per time step until stopping, and let $W(m, m')$ be the loss of making decision $D = m'$ when the true state is $\mathbf{P}_m$. It is standard to assume that $W(m, m) = 0$ but $W(m, m') > 0$ for any $m \neq m'$, i.e., no loss occurs if and only if a correct decision is made. Then when the true state of nature is $\mathbf{P}_m$, the total expected cost of a decentralized test δ is

$$\mathcal{R}_c(\delta; m) = c\mathbf{E}_m(N) + \sum_{m'} W(m, m')\mathbf{P}_m\{D = m'\}$$

where $\mathbf{E}_m$ is the expectation operator under $\mathbf{P}_m$. Hence, the Bayes risk of the decentralized test δ is

$$\mathcal{R}_c(\delta) = \sum \pi_m \mathcal{R}_c(\delta; m). \tag{17}$$

In our research, we develop asymptotically Bayes procedures by introducing a class of "two-stage" decentralized sequential tests in which each local sensor uses two stationary (possibly randomized) local quantizers with at most one switch between these two quantizers. This type of tests are useful because they allows the fusion center to first make a preliminary guess about the true state of nature and then optimize the procedure accordingly.

Our proposed two-stage test $\delta(c)$ can be defined as follows. In the *first stage* of $\delta(c)$, the local sensor can use any "reasonable" stationary deterministic quantizer and the fusion center needs to make a preliminary guess about the true state of nature. The only requirement is that as the cost $c \to 0$, the probabilities of making incorrect preliminary guess go to zero but the time steps taken at this first stage become negligible as compared to those of the overall procedure (or the second stage).

To be more concrete, let $u(c) \in (0, 1/2)$ be a function of c such that $u(c) \to 0$ and $\log u(c)/\log c \to 0$ when $c \to 0$, e.g., $u(c) = 1/|\log c|$. Choose a deterministic quantizer ϕ^0 such that $I(m, m'; \phi^0) > 0$ for any two states $0 \leq m \neq m' \leq M-1$, and let the local sensor use the stationary quantizer ϕ^0 to send i.i.d. sensor messages $U_n = \phi^0(X_n)$ to the fusion center. Then the fusion center faces a classical sequential detection problem with the i.i.d. sensor messages U_n's as inputs, and thus it is intuitively appealing to make a preliminary decision based on posterior distributions. Specifically, at each time step $n = 0, 1, \cdots$, the fusion center updates recursively the posterior distribution $(\pi_{0,n}, \pi_{1,n}, \ldots, \pi_{M-1,n})$ as follows:

$$\pi_{m,n} = \frac{\pi_{m,n-1} f_m(U_n; \phi^0)}{\sum_{0 \leq m' \leq M-1} \pi_{m',n-1} f_{m'}(U_n; \phi^0)}.$$

Then the fusion center will stop the first stage at time step

$$N_0 = \min\{n \geq 0 : \max_{0 \leq m \leq M-1} \{\pi_{m,n}\} \geq 1 - u(c)\}$$

and when stopped, the fusion center makes a preliminary decision

$$D_0 = arg \max_{0 \le m \le M-1} \pi_{m,N_0}.$$

Note that the preliminary decision D_0 is well-defined because the maximum value of π_{m,N_0} is attained at only one index m due to the definition of N_0 and the fact that $u(c) < 1/2$. For the purpose of practical implementation, the preliminary decision D_0 can be transmitted to the local sensor through a feedback of $\log_2 M$ bits.

In the *second stage* of our proposed test $\delta(c)$, the local sensor will switch to another stationary (likely randomized) quantizer that may depend on the preliminary decision D_0. Without loss of generality, we assume that the local sensor uses the stationary quantizer $\bar{\phi}_m$ when the preliminary decision at the first stage is $D_0 = m$ for $m = 0, 1, \ldots, M-1$. Here we put a bar over $\bar{\phi}_m$ to emphasize that it is likely a randomized quantizer when optimized, and we will postpone the detailed discussion about how to implement randomized quantizers to the next subsection.

Now at the second stage, the fusion center shall ignore the preliminary decision D_0 and *continue to update* the posterior distribution $(\pi_{0,n}, \ldots, \pi_{M-1,n})$ with the sensor messages generated from the new quantizer $\bar{\phi}_m$ when $D_0 = m$ (how to update will be discussed in the next subsection). Then the fusion center will stop the second stage (hence the whole procedure) at time step

$$N = \min\{n \ge N_0 : \max_{0 \le m \le M-1} \{\pi_{m,n}\} \ge 1 - c\}$$

and when stopped, the fusion center makes a final decision

$$D = arg \max_{0 \le m \le M-1} \pi_{m,N}.$$

In Wang and Mei (2010), the asymptotic optimality properties of our proposed test are established. During the process, a spin-off project is to investigate the effect of quantization on the second or higher order moments of log-likelihood ratios. The well-known Kullback-Leibler's inequality states that quantization cannot increase Kullback-Leibler information number which is just the first moment of of log-likelihood ratios. In our latest manuscript, Wang and Mei (2010) (see appendix 4.2), we extend the Kullback-Leibler's inequality to show that quantization may result in an increase of the second or higher order moments of log-likelihood ratios, but such an increase is bounded by a universal constant that only depends on the value of the moment (such a constant is $2/e$ for the second moment). Furthermore, we also illustrate that the result can be used to provide a simpler sufficient condition for the asymptotic optimality theory in decentralized sequential detection problems.

2.4 The 2-CUSUM Model for Two-Choice Reaction Time

The mathematical models have played an important roles in psychology and behavior sciences, and one important example is Ratcliff's diffusion model that has been successfully applied to account for speed-accuracy relations, mean response times and the shapes of reaction-time distributions for the paradigm of two alternative forced choice, i.e., when the subject in the experiment has to choose one of two decisions, e.g., "Signal is in Position 1" or "Signal is in Position 2." However, when the subject in the two-choice experiment is allowed to say "No signal," it is unclear how to "best" extend Ratcliff's diffusion model so that the corresponding mathematical model can take into account of uncertainty about the time of occurrence of the signal.

In our research, spurred by the experiments with uncertainty about the time of occurrence of the signal, we extend Ratcliff's diffusion model to the 2-CUSUM model by emphasizing natural connections to the sequential change-point detection problems. The proposed 2-CUSUM model is based on the optimal procedure in the problem of detecting a change in the mean of Brownian motion from $\mu = 0$ (no signals) to $\mu = \pm 2\lambda$ (a signal appears).

To state more rigorously, in Ratcliff's diffusion model, it is assumed that each subject in an experiment has an internal (real-valued) process that summarizes the information accumulation over time. The process runs between two thresholds and is terminated as soon as one of the thresholds is crossed, and the subject will choose one of two decisions depending on which threshold is crossed. In order for the model to be flexible to describe the data in real-world applications, several parameters in the internal process are assumed to be random effects to account for intra-trial variability or subject-specific effects. The key mathematical tool is based on the optimality of Wald's sequential probability ratio tests (SPRT). Mathematically, in the diffusion model, the response time (RT) is given by

$$RT = d + T_{er},$$

where $T_{er} \sim \text{Uniform}(S_t)$ and the decision time d is defined as a SPRT-type stopping time

$$d = \inf\{t \geq 0 : z + Z_t \notin (0, a)\}, \tag{18}$$

where $Z_t = vt + sW_t$, $\{W_t; t \geq 0\}$ is a standard Brownian motion, the starting point $z \sim \text{Uniform}(S_z)$, the standard deviation s is a constant, and the drift rate $v \sim N(v_0, \eta^2)$. Here the drift rate v denotes the rate of accumulation of information and it is determined by the quality of the information extracted from the stimulus. In an experiment, the value of drift rate would be different for each stimulus condition that differed in difficulty.

Our proposed 2-CUSUM process model is as follows. For a fixed $\lambda \geq 0$, for any $t \geq 0$, consider two process

$$R_{0,t} = -\lambda t + Z_t + (z - \lambda) \qquad \text{and} \qquad R_{1,t} = -\lambda t - Z_t + (a - z - \lambda). \tag{19}$$

For $l = 1, 2$, define the corresponding CUSUM processes

$$W_{l,t} = R_{l,t} - \min(0, \inf_{0 \leq s \leq t} R_{l,s}), \tag{20}$$

and the Page's CUSUM procedures

$$d_l^* = \inf\{t \geq 0 : W_{l,t} \geq a\}.$$

Then the decision time is defined as

$$d^* = \min(d_0^*, d_1^*). \tag{21}$$

There are two new features in the proposed 2-CUSUM process models. The first one is *information discounting* in the sense of adding a negative term $-\lambda t$ to both cumulative information processes, $R_{0,t}$ and $R_{1,t}$. The second is *memory forgetting* in the sense that the decision makers will "forget" or "ignore" those "not so significant" information that is not consistent with their prior information, and will start afresh to cumulative information until there is sufficient evident to prove that the prior information is incorrect. Moreover, the discounting information rate λ can also be thought of the detection limit for individuals, i.e., human being essentially ignore those information whose rate is less than λ in their decision making statistics $W_{l,t}$'s. In Moutakides and Mei (2010) (see appendix 4.3), we investigate the properties of d^* in (21), and show that d^* is equivalent to d in (18) when $\lambda = 0$. Furthermore, when $\lambda > 0$, d^* is asymptotically equivalent to d (in the sense that both are asymptotically normally distributed) when the drift rate $v \notin [-\lambda, \lambda]$, but they have completely distributions when the drift rate $v \in [-\lambda, \lambda]$.

3 Publications

1. Y. Mei. Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, vol. 97, page 419–433, 2010.

2. Y. Mei, S. W. Han and K. Tsui. Early detection of a change in Poisson rate after accounting for population size effects. To appear in *Statistica Sinica*, vol. 21, 2011.

3. Y. Mei. Discussion on 'Question detection problems: fifty years later' by Professor Shiryaev. *Sequential Analysis*, vol 29, page 410-414, 2010.

4. Y. Wang and Y. Mei. Decentralized two-sided sequential test for a normal mean. *Proceedings 2009 IEEE International Symposium on Information Theory (ISIT 2009)*, Seoul, Korea, June 28-July 3, 2009.

5. Y. Wang and Y. Mei. Asymptotic Optimality Theory For Decentralized Sequential Multihypothesis Testing Problems. Submitted to *IEEE Trans. Inform. Theory*, 2010. Available online at `http://arxiv.org/abs/1011.0228`

6. Y. Mei. Monitoring A Large Number of Data Streams Via Thresholding. 2011 (see appendix 4.1).

7. Y. Wang and Y. Mei. A Generalization of Kullback-Leibler's Inequality and Its Applications to Quantization Effects on Detection. 2010 (see appendix 4.2).

8. G. Moustakides and Y. Mei. The 2-CUSUM Test, 2010 (see appendix 4.3).

4 Appendix: Copies of (3) manuscripts under AFOSR grant

4.1 Manuscript A: Monitoring A Large Number of Data Streams Via Thresholding

MONITORING A LARGE NUMBER OF DATA STREAMS VIA THRESHOLDING

By Yajun Mei[*],

Georgia Institute of Technology

The sequential change-point detection problem is studied in a general context of monitoring a large number of data streams when the trigger event may affect different data streams differently, e.g., the subset of affected data stream is unknown and the event could have an immediate or delayed impact on affected data streams. Motivated by the scenario when one has a prior knowledge that at most r data streams will be affected and by the applications in censoring sensor networks, we propose scalable global monitoring schemes based on the sum of those local CUSUM statistics that are "large" under either hard thresholding or top-r thresholding rules or both. The proposed schemes are shown to possess certain asymptotic optimality properties in the simplest model, and extensions to more complicated models are discussed.

1. Introduction. Sequential change-point detection, or more generally, sequential methodologies, have a variety of applications such as industrial quality control, signal detection and biosurveillance. The classical version of this problem, where one monitors a single *univariate* independent and identically distributed (i.i.d.) data stream, is a well-developed area, and many classical schemes have been developed such as the Shewhart's chart

[*] Supported in part by NSF grant CCF-0830472 and AFOSR grant FA9550-08-1-0376.

AMS 2000 subject classifications: Primary 62L15, 60G40, 62F12

Keywords and phrases: change-point, CUSUM, optimal stopping, sensor networks, sequential detection

(Shewhart [24]), moving average control charts, Page's CUSUM procedure (Page [18]), Shiryaev-Roberts procedure (Shiryaev [25], Roberts [23]) and scan statistics. All these classical schemes not only hold attractive theoretical properties, but also are computationally simple. See, for example, Lorden [11], Pollak [19, 20], Moustakides [16], Lai [9, 10], Kulldorff [8].

In the modern information age, one often faces the need to monitor a large number of data streams with the aim of offering the potential for early detection of a "trigger" event. In the literature, two special categories of global monitoring have gained a lot of attention: one is consensus detection in which the occurring event is assumed to affect all data streams abruptly and simultaneously, and the other is the case when one and only one data stream is known to be affected by the occurring event, e.g., multi-channel detection in signal detection. See, for example, Montgomery [15], Veeravalli [31], Tartakovsky and Veeravalli [28] and Tartakovsky et al. [29].

In this article, we consider a general scenario when the occurring event may affect different data streams differently. For instance, we may know that the event can affect 10 out of 100 data streams but we do not know which 10 data streams will be affected. In addition, the event could have an immediate or delayed impact on affected data streams. One example is biosurveillance when a disease outbreak may first occur in some unknown local cities or counties and then spread out to other regions. Another example occurs in target detection when one uses several different sensor types such as radar, sonar, infrared and magnetic to detect a target, and it is possible that some (but not all) sensors can provide earlier information about the target.

Ideally, one would like to develop a single global monitoring scheme that can detect the occurring event as quickly as possible while controlling the system-wise global false alarm rate. While such global monitoring schemes can be found by the standard statistical methods such as generalized likelihood ratios or mixture likelihood ratios from the theoretical viewpoint, they are generally infeasible from the computational viewpoint since they would require us to do an exhaustive search over all possible post-change hypotheses or over all possible combinations of affected data streams. Hence, it is desirable to find a global monitoring scheme that is not only efficient, but also *scalable* in the sense of being able to be implemented to monitor a large number of data streams over a long period of time.

To develop a scalable global scheme, one intuitively appealing approach is to monitor each local data stream locally through some classical computationally efficient schemes and then combine all local schemes together to produce a single global scheme. Unfortunately, little research has been done on how to best combine all local schemes together to ensure efficiency, and to the best of our knowledge, only two existing approaches have been proposed so far. The first one is a naive approach that raises an alarm at the global level whenever any local scheme raises a local alarm. This naive approach is very effective if one or very few data streams are affected, but it may lose efficiency when several or more data streams can provide information to the occurring event. For the purpose of comparison, below the global scheme corresponding to this naive approach will be referred as a “MAX” scheme, since it can be thought of raising a global alarm if the *maximum* of local monitoring statistics is too large, where, if necessary, one can normalize the local schemes so that the local detection thresholds are the same. Recently, in Mei [14], the present author proposes to raise an alarm at the global level if the *sum* of local monitoring statistics (e.g., local CUSUM statistics in the logarithm scale) is too large, and theoretical analysis and numerical simulations show that the corresponding “SUM” scheme is efficient when the number of affected data streams is very large.

The objective of this article is to illustrate that there are other simple approaches to combine all local schemes together to produce an efficient scalable global scheme. Our proposed methodologies are motivated by the following two applications. The first one is the scenario in which one has a prior knowledge that at most r out of K data streams will be affected by the occurring event, especially when r is neither too small nor too large, e.g., $r = 10$ and $K = 100$. This scenario may be defined by the network fault-tolerant design to avoid risking failure. Unfortunately, the existing "MAX" or "SUM" schemes can be ineffective under this scenario, and new methodologies are needed to take advantage of such a prior knowledge in globally monitoring.

The second motivation of our proposed methodologies is censoring sensor networks in engineering, which was introduced by Rago et al. [22] and later by Appadwedula et al. [1] and by Tay et al. [30]. Figure 1 illustrates the general setting of a widely used configuration of censoring sensor networks, in which the data streams $X_{k,n}$'s are observed at the remote sensors (typically low-cost battery-powered devices), but the final decision is made at a central location, called the fusion center. The key feature of such a network is that while sensing (i.e., taking observations at the local sensors) are generally cheap and affordable, communication between remote sensors and fusion center are expensive in terms of both energy and limited bandwidth. Thus, to prolong the reliability and lifetime of the network system, practitioners often allow the local sensors to send summary messages $U_{k,n}$'s to the fusion center only when necessary. The question then becomes when and how to send summary messages so that the fusion center can still monitor the network system effectively.

The above two considerations motivate us to propose to raise a global

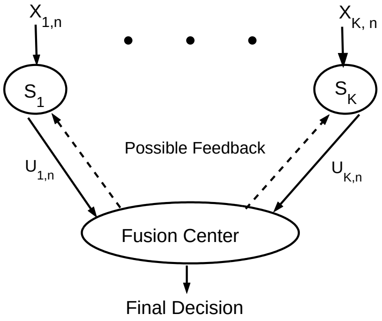

FIG 1. *General setting of a widely used configuration of censoring sensor networks.*

alarm based on the sum of those local detection statistics (e.g., local CUSUM statistics) that are "large" under either top-r thresholding rules or hard thresholding or both. It turns out that the corresponding global schemes include both MAX and SUM schemes as special cases and possess certain asymptotic optimality properties under proper criteria. It is worth pointing out that a well-known view in the standard off-line statistical inference literature is the necessity of thresholding for high-dimensional data in order to improve power or efficiency. Thus, from the methodology point of view, our proposed methodologies are analogous to those off-line statistical methods such as (adaptive) truncation, and soft and hard thresholding, see Neyman [17], Donoho and Johnstone [4], Fan and Lin [6]. Also see Candes [3] and the references there. However, our motivation here is different and our application to sequential change-point detection is new.

The remainder of this article is organized as follows. In Section 2 we present a rigorous mathematical formulation of sequential change-point detection problems in the context of globally monitoring multiple data streams and also discuss existing methodologies. Our proposed methodologies are defined in Section 3 for the simplest model. Section 4 reports numerical Monte Carlo simulation results, and Section 5 presents an asymptotic optimality theory. Section 6 extends our methodologies to more complicated models. The appendices include the proofs of our main theorems, Theorems 5.1 and 5.2.

2. Problem Formulation and Existing Methodologies. To illustrate our main ideas, we begin with the simplest independent model in which the pre-change and post-change distributions are completely specified, and more complicated models will be discussed in Section 6. To be more specific, assume that in a system, one is monitoring K data streams over time n, say, $\{X_{k,n}\}_{n=1}^{\infty}$ for $k = 1, \ldots, K$. Initially, the system is "in control" and the distribution of the $X_{k,n}$'s is f_k for the k-th data stream. At some *unknown* time ν, a "trigger" event occurs and affects the system in the sense that the density function of the observations $X_{k,n}$'s changes from f_k to another given density g_k at time $\nu_k \geq \nu$. Without loss of generality, we assume that $\min_{1 \leq k \leq K} \nu_k = \nu$, since otherwise the system is affected by the event only after the time $\nu' = \min_{1 \leq k \leq K} \nu_k$, which can be treated as the new change-point. It is desired to utilize the observed data streams $X_{k,n}$'s to raise an alarm at the global level as soon as the event occurs so that one can take appropriate action.

A standard minimax formulation is to minimize the detection delay subject to a false alarm constraint. The latter is typical of the form

$$\mathbf{E}^{(\infty)}(T) \geq \gamma, \tag{2.1}$$

where T is the stopping rule associated with the detection scheme, $\gamma > 0$ is a pre-specified constant, and $\mathbf{E}^{(\infty)}$ denote the expectation when there are no changes. In the literature $\mathbf{E}^{(\infty)}(T)$ is often called as the average run length to false alarm. The definition of detection delay is a little more complicated, as it needs to take into account the uncertainty of the change-point ν. One widely used definition is the following "worst case" detection delay defined in Lorden [11],

$$\mathbf{D}(T) = \sup_{1 \le \nu < \infty} \operatorname{ess\,sup} \mathbf{E}^{(\nu)} \Big((T - \nu + 1)^{+} \Big| \mathcal{F}_{\nu-1} \Big), \tag{2.2}$$

where $x^{+} = \max(x, 0)$, $\mathcal{F}_{\nu-1} = (X_{1,[1,\nu-1]}, \ldots, X_{K,[1,\nu-1]})$ denotes past global information at time ν, and $X_{k,[1,\nu-1]} = (X_{k,1}, \ldots, X_{k,\nu-1})$ is past local information for the k-th data stream. Note that $\mathbf{E}^{(\nu)}$ in the definition (2.2) is obscure as it does not takes into account different ν_k's for different data streams, see (5.2) below for a more precise definition. It is often but not always that the worst case detection delay of a scheme is attained when the change-point $\nu = 1$, since it is generally more difficult to detect when a change occurs at earlier stages rather than at latter stages.

Before discussing the existing schemes for globally monitoring, let us consider the local schemes for monitoring a single data stream, say, the kth data stream. Such a problem has been well studied in the sequential change-point detection literature, and one efficient local scheme is Page's CUSUM procedure which raise a local alarm as soon as the local CUSUM statistic exceeds some pre-specified threshold, where the local CUSUM statistic for the kth data stream at time n is given by

$$W_{k,n} = \max \Big\{ 0, \max_{1 \le \nu \le n} \sum_{i=\nu}^{n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \Big\}.$$

In practice, the local CUSUM statistic $W_{k,n}$ can be computed conveniently online via a recursive formula:

$$W_{k,n} = \max \Big(W_{k,n-1} + \log \frac{g_k(X_{k,n})}{f_k(X_{k,n})}, \ 0 \Big), \tag{2.3}$$

and $W_{k,0} = 0$.

Now let us go back our global monitoring problem. As mentioned in the introduction, two approaches have been proposed to combine all local schemes together to develop a single scalable global scheme. The first one is the "MAX" scheme that raises an alarm at the global level as soon as one of local CUSUM procedures raises a local alarm. Mathematically, the "MAX" scheme raises a global alarm at time

$$(2.4) \qquad T_{\max}(c) = \text{ first time } n \geq 1 \text{ such that } \max_{1 \leq k \leq K} W_{k,n} \geq c,$$

($= \infty$ if such n does not exist) where $c > 0$ is a pre-specified constant chosen to satisfy the false alarm constraint (2.1). The second approach is the "SUM" scheme, proposed in Mei [14], in which one raises an alarm when the sum of local CUSUM statistics is too large. Specifically, at time n, each data streams calculates its local CUSUM statistic $W_{k,n}$'s as in (2.3), and then one will raise an alarm at the global level at time

$$(2.5) \qquad T_{\text{sum}}(d) = \text{ first time } n \geq 1 \text{ such that } \sum_{k=1}^{K} W_{k,n} \geq d,$$

($= \infty$ if such n does not exist) where the constant $d > 0$ is some suitably chosen constant. The intuition behind the "SUM" rule is that a large value of local CUSUM statistic indicates a possible local change, and thus the sum of local CUSUM statistics is a reasonable detection statistic if the number of affected data streams is large or completely unknown. Intuitively, the "MAX" scheme $T_{\max}(c)$ in (2.4) works better when one or very few data streams are affected, whereas the "SUM" scheme $T_{\text{sum}}(d)$ in (2.5) works better when many data streams are affected, and numerical simulations in Mei [14] indeed verified this intuition.

3. The Proposed Methodologies.

In this section, we propose to develop global monitoring schemes via top-r thresholding and/or hard thresholding. It turns out that the proposed schemes integrate the existing "MAX" and "SUM" schemes, and offer new methodologies for online monitoring a large number of data streams.

3.1. *Top-r Thresholding Schemes.* In some applications, one may have a prior knowledge that at most r out of K data streams will be affected by the occurring event. This inspires us to construct the global monitoring schemes based on the largest r local CUSUM statistics. To be more concrete, at time n, order the K local CUSUM statistics $W_{1,n}, \ldots, W_{K,n}$ from largest to smallest: $W_{(1),n} \geq W_{(2),n} \geq \ldots \geq W_{(K),n}$, and then the top-$r$ thresholding scheme raises a global alarm at time

$$N_{top,r}(a) = \text{ first time } n \geq 1 \text{ such that } \sum_{k=1}^{r} W_{(k),n} \geq a. \tag{3.1}$$

Obviously, the top-r thresholding scheme $N_{top,r}(a)$ becomes the "MAX" scheme $T_{\max}(c)$ when $r = 1$, and becomes the "SUM" scheme $T_{\text{sum}}(d)$ when $r = K$. Intuitively, when the occurring event affects (at most) r out of K data streams, then (at most) r local CUSUM statistics will be significantly large to provide information to the occurring event, and thus one should expect $N_{top,r}(a)$ to be suitable and effective to detect such a scenario.

3.2. *Hard Thresholding Schemes.* In censoring sensor networks illustrated in Figure 1, the local sensors need to summarize the information and only send significant information to the fusion center to prolong the reliability and lifetime of the network system. This inspires us to propose to transmit only those local CUSUM statistics $W_{k,n}$'s that are larger than their respective local thresholds, and the corresponding scheme will be called hard thresholding schemes.

Specifically, at time n, each local sensor (data stream) calculates its local CUSUM statistic $W_{k,n}$ recursively as in (2.3). Then at time n, the sensor message from the sensor to the fusion center is given by

$$U_{k,n} = \begin{cases} W_{k,n}, & \text{if } W_{k,n} \geq b_k \\ \text{NULL}, & \text{if } W_{k,n} < b_k \end{cases},$$

where $b_k \geq 0$ is the local censoring (hard threshold) parameter at the k-th sensor (or data stream). Here the message "NULL" is a special sensor symbol to indicate the local CUSUM statistic is not large. In practice, "NULL" could be represented by the situation when the sensor does not send any messages to the fusion center, e.g., the sensor is silent.

After receiving the local sensor messages $U_{k,n}$'s from the sensors, the fusion center then raises an alarm at the global level at first time n such that

$$\sum_{\text{received}} U_{k,n} \geq a,$$

where the detection threshold $a > 0$ is a pre-specified constant.

So far we simply follow our intuition without discussing how to choose the local censoring parameters b_k's, especially when the data streams are nonhomogeneous. It turns out that a "good" choice of the b_k's is $b_k = \rho_k b$ for $k = 1, \ldots, K$ for some constant $b \geq 0$ where

$$\rho_k = \frac{I(g_k, f_k)}{\sum_{k=1}^{K} I(g_k, f_k)} \tag{3.2}$$

can be thought of as the weight of the k-th data stream in the overall final decision, and $I(g_k, f_k)$ is the Kullback-Leibler information number defined by

$$I(g_k, f_k) = \int \log \frac{g_k(x)}{f_k(x)} g_k(x) d\mu(x). \tag{3.3}$$

With this choice of b_k's, the proposed hard thresholding scheme will raise a global alarm at time

$$N_{hard}(a, b) = \text{ first } n \geq 1 \text{ such that } \sum_{k=1}^{K} W_{k,n} I\{W_{k,n} \geq \rho_k b\} \geq a, \tag{3.4}$$

where $a > 0$ and $b \geq 0$ are two suitably chosen constants.

Evidently, if the threshold parameter $b = 0$, then the hard thresholding scheme $N_{hard}(a, b)$ in (3.4) becomes the "SUM" scheme $T_{\text{sum}}(d)$ in (2.5) since the $W_{k,n}$'s are always non-negative by the definition in (2.3). On the other hand, if the threshold parameter b is very large, say $b \geq a / \min_{1 \leq k \leq K} \rho_k$, then the hard thresholding scheme $N_{hard}(a, b)$ in (3.4) becomes the "MAX" scheme $T_{\max}(c)$ in (2.4) with $c = \min_k(\rho_k b)$. Therefore, the family of schemes $N_{hard}(a, b)$ is actually a very large family that includes both "MAX" and "SUM" schemes.

It is interesting to note that the proposed hard thresholding and top-r thresholding schemes are closely related to the concept of "censoring" in engineering, statistics, and medical science, particularly reliability and survival analysis. Specifically, the hard thresholding scheme is in parallel to the so-called "Type I censoring" in which an experimenter has a set number of subjects or items and stops the experiment at a pre-determined time, whereas the top-r thresholding scheme is in parallel to the so-called "Type II censoring" in which one stops the experiment when a predetermined number of subjects are observed to have certain properties. In addition, the top-r thresholding schemes can also be thought of hard-thresholding rules where the local censoring parameters are data-driven and adaptive over time n. Specifically, let $w_{r,n}$ denote the r-th largest statistics, i.e., $w_{r,n} = W_{(r),n}$, and then one raises an alarm at the global level at the time

$$N^*_{top,r}(a) = \text{ first } n \geq 1 \text{ such that } \sum_{k=1}^{K} W_{k,n} I\{W_{k,n} \geq w_{r,n}\} \geq a.$$

Rigorously speaking, we have $N_{top,r}(a) \leq N^*_{top,r}(a)$, and they are equivalent only when the $W_{k,n}$'s are non-arithmetic, since they can be different when more than one of $W_{k,n}$'s is equal to $w_{r,n}$. Fortunately, both $N_{top,r}(a)$ and

$N^*_{top,r}(a)$ possess similar asymptotic optimality properties and either can be used in our context.

3.3. *The Combined Thresholding Schemes.* For the purpose of applications in censoring sensor networks in Figure 1, one may combine the abovementioned two thresholding methods together: use the hard thresholding at the local sensor level, and then adopt the top-r thresholding rule at the fusion center level to detect the event when one has a prior knowledge that the occurring event affects at most r sensors. Specifically, at time n, the sensor message sent by the k-th local sensor is defined by $U_{k,n} = W_{k,n} I\{W_{k,n} \geq \rho_k b\}$. The fusion center then orders all sensor messages $U_{k,n}$'s as $U_{(1),n} \geq \dots \geq U_{(K),n}$, and raises an alarm if the sum of the r largest $U_{k,n}$'s is too large. Mathematically, the combined thresholding scheme raises an alarm at the global level at time

$$N_{comb,r}(a,b) = \text{ first } n \geq 1 \text{ such that } \sum_{k=1}^{r} U_{(k),n} \geq a. \tag{3.5}$$

Alternatively, let $u_{r,n}$ be the r-th largest value among the K sensor messages $U_{k,n}$'s, and then another version of the combined thresholding scheme can be defined by the stopping time

$$\begin{aligned} N^*_{comb,r}(a,b) &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} U_{k,n} I\{U_{k,n} \geq u_{r,n}\} \geq a\Big\} \\ &= \inf\Big\{n \geq 1 : \sum_{k=1}^{K} [W_{k,n} I\{W_{k,n} \geq \rho_k b\} I\{W_{k,n} \geq u_{r,n}\}] \geq a\Big\}. \end{aligned}$$

Again, $N_{comb,r}(a,b) \leq N^*_{comb,r}(a,b)$, but they are equivalent in non-arithmetic cases and are generally asymptotically equivalent otherwise. Evidently, the family of the combined thresholding schemes contains two censoring parameters, b and r, and it includes the families of both hard-thresholding and top-r thresholding schemes.

3.4. *Choice Of Thresholding Parameters.* Compared to the existing "MAX" or "SUM" schemes, our proposed thresholding schemes include two new thresholding parameters: r for the top-r thresholding schemes and b for the hard-thresholding schemes (and both r and b for the combined thresholding schemes). It is natural to ask how to choose these two thresholding parameters in practice?

The choice of thresholding parameter r is straightforward and depends on whether one has any prior knowledge about the maximum number of affected data streams. If such a knowledge exists and it is believed that at most r_0 data streams will be affected by the occuring event, then one should use this r_0 as the value of thresholding parameter r. Otherwise one may want to be conservative to choose $r = K$, e.g., consider the "SUM" scheme or the hard-thresholding scheme $N_{hard}(a, b)$ in (3.4).

The choice of thresholding parameter b is nontrivial, and may need to consider some non-statistical constraints. As an illustration, in certain applications of censoring sensor networks, the censoring parameter b may be chosen to satisfy the constraints on the average fraction of transmitting sensors when no events occur. For our proposed scheme $N_{hard}(a, b)$, when no event occurs, the average fraction of transmitting sensors at any time step n is

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(U_{k,n} \neq \text{NULL}) &= \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}^{(\infty)}(W_{k,n} \geq \rho_k b) \\ &\leq \frac{1}{K}\sum_{k=1}^{K} \exp(-\rho_k b), \end{aligned}$$

where the last inequality follows from the well-known properties of the local CUSUM statistics that $\mathbf{P}^{(\infty)}(W_n \geq a) \leq \exp(-a)$ for all $a > 0$, see, for example, Appendix 2 on Page 245 of Siegmund [26]. In particular, if all K sensors are homogeneous in the sense that the $I(g_k, f_k)$'s are the same for all k, then $\rho_k = 1/K$, and the average fraction of transmitting sensors at any time step is $\exp(-b/K)$ when no event occurs. Hence for our proposed scheme $N_{hard}(a, b)$, a choice of

$$b = K \log \eta^{-1},$$

or equivalently, the local hard threshold $b_k = \rho_k b = b/K = \log \eta^{-1}$, will guarantee that on average, at most $100\eta\%$ of K homogeneous sensors will transmit messages at any given time when no event occurs. It is worth emphasizing that here the thresholding parameter b is a function of K and the local threshold $b_k = \log \eta^{-1}$ is a constant that does not depend on K.

The choice of b becomes more complicated for the combined thresholding schemes $N_{comb,r}(a, b)$ (or $N^{*}_{comb,r}(a, b)$) if the thresholding parameter r has been given beforehand. We do not have an explicit answer, and a general rule of thumb that the censoring parameter b in (3.5) shall not be too large, as one generally should keep at least r non-zero $U_{k,n}$'s when r data streams are affected by the event.

4. Numerical Simulations.

Before we offer asymptotic optimality theory, let us present a numerical simulation study in this section to illustrate the usefulness of the proposed schemes. Suppose that there are $K = 100$ independent and identical sensors in a system illustrated in Figure 1, and each local sensor observes a one-dimensional data stream. Assume that the observations at each local sensor are i.i.d. with mean 0 and variance 1 before the change and with mean 0.5 and variance 1 after the possible change. In our simulation study, we simply assume that the change is instantaneous if a sensor is affected, i.e., the local change-point $\nu_k = \nu$ or ∞, and we do not know which subset of sensors will be affected by the occurring event.

For the purpose of comparison, we conduct numerical simulations for five families of detection schemes:

- the "MAX" scheme $T_{\max}(c)$ in (2.4),
- the "SUM" scheme $T_{\text{sum}}(d)$ in (2.5),
- the top-r thresholding scheme $N_{top,r}(a)$ in (3.1) with $r = 10$,
- the hard thresholding scheme $N_{hard}(a, b)$ in (3.4),
- the combined thresholding schemes $N_{comb,r}(a, b)$ in (3.5) with $r = 10$.

For each family of schemes $N_{hard}(a, b)$ and $N_{comb,r=10}(a, b)$, we further consider three specific schemes, depending on the value of the hard-thresholding parameter b: (i) $b = K/2 \approx -\log(0.607) * K$, (ii) $b = -\log(0.1) * K = 2.3026K$ and (iii) $b = -\log(0.01) * K = 4.6052K$. In the context of censoring sensor networks, the choices of these values will guarantee that when no event occurs, on average at most $\eta = 60.7\%, 10\%$, and 1% of $K = 100$ homogeneous sensors will transmit messages at any given time, respectively. Hence, in our simulation study, there are a total of nine specific monitoring schemes.

In order to fairly compare these nine specific schemes $T(a)$, we applied them to the same computer-generated pseudo-random data sets in our simulation studies. Specifically, we first simulate a single large random matrix with dimension $m \times nmax \times K$ from the pre-change distributions, where m denotes the desired number of repetitions in Monte Carlo simulations, $nmax$ is the largest time step (or the largest run length to raise an alarm), and K is the number of data streams (or sensors). In our numerical simulations, $m = 1000, nmax = 2 * 10^5$ and $K = 100$. Then, for each of these nine specific schemes $T(a)$, the simulated data set was used to determine the appropriate values of the detection threshold a to satisfy $\mathbf{E}_\infty(T(a)) \approx \gamma$ (within the

Table 1

Detection delays with $K = 100$ identical data streams: 20 or more streams are affected

	Detection	# affected data streams				
γ	Scheme	100	80	50	30	20
10^3	$T_{\text{sum}}(d = 101.66)$	5.55 ± 0.02	6.53 ± 0.02	9.14 ± 0.04	13.0 ± 0.1	17.3 ± 0.1
	$N_{hard}(a = 96.79, b = 50)$	5.58 ± 0.02	6.54 ± 0.02	9.16 ± 0.04	13.0 ± 0.1	17.4 ± 0.1
	$N_{hard}(a = 52.28, b = 230.26)$	7.72 ± 0.02	8.49 ± 0.03	10.47 ± 0.04	13.6 ± 0.1	17.1 ± 0.1
	$N_{hard}(a = 21.90, b = 460.52)$	12.57 ± 0.04	13.30 ± 0.05	14.91 ± 0.06	17.1 ± 0.1	19.5 ± 0.1
	$T_{\max}(c = 8.77)$	22.46 ± 0.11	23.23 ± 0.11	24.64 ± 0.13	26.9 ± 0.2	28.7 ± 0.2
	Schemes $N_{comb,r}(a, b)$ in (3.5)					
	$N_{top,r=10}(a = 41.67)$	10.82 ± 0.03	11.48 ± 0.04	13.07 ± 0.05	15.3 ± 0.1	17.9 ± 0.1
	$N_{comb,r=10}(a = 41.67, b = 50)$	10.82 ± 0.03	11.48 ± 0.04	13.07 ± 0.05	15.3 ± 0.1	17.9 ± 0.1
	$N_{comb,r=10}(a = 41.49, b = 230.26)$	10.73 ± 0.03	11.41 ± 0.04	12.98 ± 0.05	15.2 ± 0.1	17.8 ± 0.1
	$N_{comb,r=10}(a = 21.90, b = 460.52)$	12.57 ± 0.04	13.30 ± 0.05	14.91 ± 0.06	17.1 ± 0.1	19.5 ± 0.1
10^4	$T_{\text{sum}}(d = 111.04)$	6.16 ± 0.02	7.29 ± 0.02	10.25 ± 0.04	14.9 ± 0.1	20.1 ± 0.1
	$N_{hard}(a = 106.38, b = 50)$	6.17 ± 0.02	7.29 ± 0.02	10.27 ± 0.04	14.9 ± 0.1	20.2 ± 0.1
	$N_{hard}(a = 62.26, b = 230.26)$	8.34 ± 0.03	9.22 ± 0.03	11.53 ± 0.04	15.3 ± 0.1	19.7 ± 0.1
	$N_{hard}(a = 29.70, b = 460.52)$	13.39 ± 0.04	14.17 ± 0.05	16.14 ± 0.06	19.0 ± 0.1	21.9 ± 0.1
	$T_{\max}(c = 11.12)$	31.79 ± 0.14	32.74 ± 0.15	34.81 ± 0.17	37.6 ± 0.2	39.9 ± 0.2
	Schemes $N_{comb,r}(a, b)$ in (3.5)					
	$N^*_{top,r=10}(a = 46.55)$	12.67 ± 0.04	13.41 ± 0.04	15.22 ± 0.05	17.8 ± 0.1	20.8 ± 0.1
	$N_{comb,r=10}(a = 46.55, b = 50)$	12.67 ± 0.04	13.41 ± 0.04	15.22 ± 0.05	17.8 ± 0.1	20.8 ± 0.1
	$N_{comb,r=10}(a = 46.53, b = 230.26)$	12.67 ± 0.04	13.41 ± 0.04	15.21 ± 0.05	17.8 ± 0.1	20.8 ± 0.1
	$N_{comb,r=10}(a = 29.70, b = 460.52)$	13.39 ± 0.04	14.17 ± 0.04	16.14 ± 0.06	19.0 ± 0.1	21.9 ± 0.2

range of sampling error based on m repetition Monte Carlo simulations). Next, using the obtained threshold value a, we simulated the detection delay when the change-point occurs at time $\nu = 1$ under several different post-change scenarios (i.e., different number of affected data streams) by adding the post-change mean $\mu_1 = 0.5$ appropriately to the simulated data matrix. This will provide the estimated values of the worst case detection delays because, for each of these nine schemes we considered, the worst case detection delay occurs when the change-point $\nu = 1$.

In our simulations we consider several post-change scenarios, depending on how many data streams are affected. To better summarize our simulations, we have divided our results into two tables: Table 1 for the cases when 20 or more data streams are affected and Table 2 for the cases when 10 or less data streams are affected. In each table, two different values of the false

Table 2

Detection delays with $K = 100$ identical data streams: 10 or less streams are affected

	Detection	# affected data streams				
γ	Scheme	10	8	5	3	1
10^3	$T_{\text{sum}}(d = 101.66)$	27.6 ± 0.2	32.5 ± 0.2	44.1 ± 0.4	61.3 ± 0.6	127.0 ± 1.7
	$N_{hard}(a = 96.79, b = 50)$	27.9 ± 0.2	32.7 ± 0.2	44.8 ± 0.4	62.1 ± 0.6	128.8 ± 1.7
	$N_{hard}(a = 52.28, b = 230.26)$	26.5 ± 0.2	30.6 ± 0.2	41.7 ± 0.3	59.4 ± 0.6	124.6 ± 1.6
	$N_{hard}(a = 21.90, b = 460.52)$	25.5 ± 0.2	28.2 ± 0.2	35.4 ± 0.3	46.9 ± 0.4	99.1 ± 1.2
	$T_{\max}(c = 8.77)$	33.0 ± 0.2	34.5 ± 0.3	38.6 ± 0.3	44.4 ± 0.4	65.8 ± 0.9
	Schemes $N_{comb,r}(a, b)$ in (3.5)					
	$N_{top,r=10}(a = 41.67)$	24.3 ± 0.2	27.1 ± 0.2	34.6 ± 0.3	45.9 ± 0.4	89.3 ± 1.1
	$N_{comb,r=10}(a = 41.67, b = 50)$	24.3 ± 0.2	27.1 ± 0.2	34.6 ± 0.3	45.9 ± 0.4	89.3 ± 1.1
	$N_{comb,r=10}(a = 41.49, b = 230.26)$	24.3 ± 0.2	27.2 ± 0.2	35.1 ± 0.3	47.1 ± 0.4	92.0 ± 1.1
	$N_{comb,r=10}(a = 21.90, b = 460.52)$	25.5 ± 0.2	28.2 ± 0.2	35.4 ± 0.3	46.9 ± 0.4	99.1 ± 1.2
10^4	$T_{\text{sum}}(d = 111.04)$	33.4 ± 0.2	39.3 ± 0.3	55.2 ± 0.4	80.2 ± 0.7	191.6 ± 2.1
	$N_{hard}(a = 106.38, b = 50)$	33.8 ± 0.2	39.8 ± 0.3	56.1 ± 0.5	81.2 ± 0.7	195.5 ± 2.1
	$N_{hard}(a = 62.26, b = 230.26)$	31.9 ± 0.2	37.7 ± 0.2	53.7 ± 0.4	78.4 ± 0.7	191.6 ± 2.1
	$N_{hard}(a = 29.70, b = 460.52)$	29.9 ± 0.2	33.5 ± 0.2	43.3 ± 0.3	61.3 ± 0.5	152.6 ± 1.7
	$T_{\max}(c = 11.12)$	45.2 ± 0.3	47.2 ± 0.3	52.3 ± 0.4	59.2 ± 0.5	85.5 ± 1.0
	Schemes $N_{comb,r}(a, b)$ in (3.5)					
	$N_{top,r=10}(a = 46.55)$	28.6 ± 0.2	32.2 ± 0.2	41.8 ± 0.3	57.1 ± 0.5	124.2 ± 1.4
	$N_{comb,r=10}(a = 46.55, b = 50)$	28.6 ± 0.2	32.2 ± 0.2	41.8 ± 0.3	57.1 ± 0.5	124.2 ± 1.4
	$N_{comb,r=10}(a = 46.53, b = 230.26)$	28.6 ± 0.2	32.3 ± 0.2	42.3 ± 0.3	58.3 ± 0.5	128.0 ± 1.4
	$N_{comb,r=10}(a = 29.70, b = 460.52)$	29.9 ± 0.2	33.5 ± 0.2	43.4 ± 0.3	61.3 ± 0.5	152.6 ± 1.8

alarm constraint γ in (2.1) are considered: $\gamma = 10^3$ and 10^4, and the detection delay is recorded as the Monte carlo estimate ± standard error. Moreover, for each given false alarm constraint γ, we further divided our results into two sub-scenarios: one for the family of schemes $N_{hard}(a, b)$ and one for the family of schemes $N_{comb,r=10}(a, b)$. Recall that under our numerical simulations setting, the "MAX" scheme $T_{\max}(c)$ can be thought of $N_{hard}(a, b)$ with $b = Kc = 100c$ and $a = c$, while the "SUM" scheme $T_{\text{sum}}(d)$ can be thought of $N_{hard}(a, b)$ with $b = 0$ and $a = d$. Likewise, the top-r thresholding scheme $N_{top,r=10}(a)$ in (3.1) can be thought of $N_{comb,r=10}(a, b)$ with $b = 0$. Thus in each simulation scenario, we report the numerical results of these specific schemes in order of increasing values of the censoring parameter b.

From Tables 1 and 2, among these nine specific schemes, when a small number ($1 \sim 3$) of 100 homogeneous data streams are affected by the event, the "MAX" scheme $T_{\max}(c)$ is the best (in the sense of smallest detection delay), the "SUM" scheme $T_{\text{sum}}(d)$ is the worst, and all other schemes $N_{hard}(a, b)$'s or $N_{comb,r}(a, b)$'s are in-between. Similarly, when a large number (20 or more) of 100 homogeneous data streams are affected, the order is reserved: $T_{\text{sum}}(d)$ is the best, $T_{\max}(c)$ is the worst, and all other schemes are in-between. However, when $5 \sim 10$ data streams are affected, the best schemes are the family of schemes $N_{comb,r=10}(a, b)$, which is designed to detect the scenario when 10 data streams are affected by the event. In addition, for each given scheme, the fewer affected data streams we have, the larger detection delay it will have. Similarly, the larger value the false alarm constraint γ in (2.1), the larger detection delay it will have. All these results are consistent with our intuition.

It worths mentioning that for the family of the hard-thresholding schemes $N_{hard}(a, b)$ in (3.4), a larger censoring threshold value b actually leads to a smaller detection delay when only a few (between 1 and 5) of data streams are affected (recall that $T_{\max}(c)$ can be thought of $N_{hard}(a, b)$ with the largest censoring parameter $b = 100c$). This is consistent with our earlier discussion that a larger censoring threshold value b in $N_{hard}(a, b)$ may actually be necessary for efficient detection when the affected data streams are sparse.

A surprising and possibly counter-intuitive result in Tables 1 and 2 is the effect of not so large values of hard-thresholding parameter b in finite sample simulations. For example, the performances of the "SUM" scheme $T_{\text{sum}}(d)$ and the hard thresholding scheme $N_{hard}(a, b = 50)$ are similar in view of sampling errors. Likewise, the top-r thresholding scheme $N_{top,r=10}(a)$ and the combined thresholding scheme $N_{comb,r=10}(a, b = 50)$ also have identical performances. In other words, for the family of schemes $N_{hard}(a, b)$ or $N_{comb,r}(a, b)$, the schemes with $b = 0$ or $b = 50$ have similar performances, even though $b = 50$ implies that the scheme only requires $\exp(-b/K) = \exp(-0.5) = 60.7\%$ of 100 sensors to transmit information to the fusion center at any given time when no event occurs.

It is also interesting to see the effect of the top-r thresholding parameter r in finite sample simulations when the hard-thresholding parameter b is large. From Tables 1 and 2, when the false alarm constraint γ in (2.1) is only moderately large, the performances of $N_{hard}(a, b = 460.52)$ and $N_{comb,r=10}(a, b = 460.52)$ are identical, and they actually also have the same detection threshold a. Intuitively, the stopping time $N_{comb,r}(a, b)$ is decreasing as a function of r, and thus we have $N_{hard}(a, b) \leq N_{comb,r=10}(a, b)$ when $b = 460.52$. So one may wonder why our numerical simulations lead to identical results? One explanation is that with such a choice of $b = 460.52$, when no event occurs, on average there is at most 1 non-zero sensor messages at any given time, and thus there is little difference whether one uses the sum of the largest $r = 10$ sensor messages or uses the sum of all $K = 100$ sensor messages. Hence similar performances are observed in finite-sample simulations.

In summary, from the performance viewpoint, using one of hard-thresholding and top-r thresholding approaches may be sufficient in certain applications, since the performance of the combined censoring scheme $N_{comb,r}(a, b)$ can be similar to that of either the hard-thresholding scheme $N_{hard}(a, b)$ or the top r-thresholding scheme $N_{top,r}(a)$, especially when the false alarm constraint γ in (2.1) is only moderately large.

5. Asymptotic Optimality Theory. In this section we establish asymptotic optimality properties of our proposed thresholding schemes. To simplify our arguments, denote by $\delta_k = \nu_k - \nu \geq 0$ the delay effect on the k-th data stream, and let $\delta_{\max} = \max_{\delta_k < \infty} \delta_k$ be the maximum delay among all finite delays. Recall that we assume $\min_k \nu_k = \nu$ and thus $\min_k \delta_k = 0$. Therefore, if we denote by Δ the set of all possible post-change hypotheses on the delay effects δ_k's and/or the subset of affected data streams, then the set Δ can be written as

$$\begin{aligned} \Delta &= \big\{(\delta_1, \ldots, \delta_K) : \text{the } \delta_k\text{'s either} = \infty \\ &\qquad \text{or satisfy } 0 \leq \delta_k \leq \delta_{\max} \text{ and } \min_{1 \leq k \leq K} \delta_k = 0\big\}. \end{aligned} \tag{5.1}$$

Moreover, we need to provide a more rigorous definition of detection delay than that in (2.2) to reflect the delay effects δ_k's. Denote by $\mathbf{P}^{(\nu)}_{\delta_1,\ldots,\delta_K}$ and $\mathbf{E}^{(\nu)}_{\delta_1,\ldots,\delta_K}$ the probability measure and expectation when the event occurs at time ν and the density of observations $X_{k,n}$'s at the k-the data stream changes from f_k to g_k at time $\nu_k = \nu + \delta_k$ for all $k = 1, \ldots, K$. As in the definition $\mathbf{D}(T)$ in (2.2), we can define the detection delay than as

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T) = \sup_{1 \leq \nu < \infty} \operatorname{ess\,sup} \mathbf{E}^{(\nu)}_{\delta_1,\ldots,\delta_K}\Big((T - \nu + 1)^+ \Big| \mathcal{F}_{\nu-1}\Big). \tag{5.2}$$

Here we use $\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T)$ to emphasize that the detection delays may depend on the delay effects δ_k's. In our asymptotic theory, we assume that the delay effects set Δ in (5.1) and the unknown change-point ν are separated, i.e., the set Δ does not depend on the unknown change-point ν. Under our setting, detecting the unknown change-point ν is of primary interest, and the delay effects δ_k's are nuisance parameters that belong to some pre-specified set Δ depending on our prior knowledge of the event.

Let us present asymptotic theory when Δ is defined in (5.1). The following theorem, whose proof is postponed in the appendix, derives the information bound on the detection delays of any globally monitoring schemes, as the false alarm constraint γ in (2.1) goes to ∞.

Theorem 5.1. *Assume a scheme $T(\gamma)$ satisfies the false alarm constraint (2.1). Then for any given post-change hypothesis $(\delta_1, \ldots, \delta_K) \in \Delta$, as γ goes to ∞,*

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T(\gamma)) \geq (1 + o(1)) \frac{\log \gamma}{J(\delta_1, \ldots, \delta_K)}, \tag{5.3}$$

where

$$J(\delta_1, \ldots, \delta_K) = \sum_{k=1}^{K} I(g_k, f_k) I\{\delta_k < \infty\}, \tag{5.4}$$

and $I(g_k, f_k)$ is the Kullback-Leibler information number defined in (3.3), and $I\{A\}$ is the indicator function of set A.

Next, we establish the asymptotic properties of the top-r thresholding scheme $N_{top,r}(a)$ in (3.1), the hard thresholding scheme $N_{hard}(a, b)$ in (3.4), and the combined thresholding scheme $N_{comb,r}(a, b)$ in (3.5) when the detection threshold a goes to ∞, regardless of the false alarm constraint (2.1). The proof of the following theorem is very technical and is presented in detail in the appendix.

Theorem 5.2. *Let $N_{a,b}$ be the proposed thresholding schemes $N_{top,r}(a)$, $N_{hard}(a, b)$ or $N_{comb,r}(a, b)$. As $a \to \infty$, let $b' = b'(a)$ be a constant such that both b' and $a - b'$ go to ∞.*

(i) *The proposed scheme $N_{a,b}$ satisfies*

$$\mathbf{E}^{(\infty)}(N_{a,b}) \geq \frac{e^a}{1 + a + \frac{a^2}{2!} + \cdots + \frac{a^{K-1}}{(K-1)!}} \tag{5.5}$$

for all $b \geq 0$.

(ii) *For any combination $(\delta_1, \ldots, \delta_K) \in \Delta$ defined in (5.1), and for all $0 \leq b \leq b'$, we have*

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{a,b}) \leq \frac{a}{J(\delta_1, \ldots, \delta_K)} + O(\sqrt{b'}) + \delta_{\max}, \tag{5.6}$$

where $J(\delta_1, \ldots, \delta_K)$ is defined in (5.4), and the relation holds for the top-r thresholding scheme $N_{top,r}(a)$ in (3.1) and the combined thresholding scheme $N_{comb,r}(a,b)$ in (3.5) with $b \geq 0$ whenever $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$, i.e., when the occurring event affects at most r data streams.

Finally, we are now in a position to establish the asymptotic optimality properties of our proposed thresholding schemes $N_{a,b}$ in the sequential change-point detection problems with the false alarm constraint γ in (2.1), as γ goes to ∞. Here we will make an additional assumption that $\delta_{\max} = o(\log \gamma)$ as γ goes to ∞, and such assumption can be easily satisfied when all finite δ_k's are uniformly bounded by some constant that does not depend on the false alarm constraint γ.

COROLLARY 5.1. *For a given K and for any $b \geq 0$, with the choice of*

$$a = a_\gamma = \log \gamma + (K - 1 + o(1)) \log\log \gamma, \tag{5.7}$$

the proposed threholding scheme $N_{a,b}$ satisfies the false alarm constraint (2.1). Moreover, if we assume that $\delta_{\max} = o(\log \gamma)$, then the hard-thresholding schemes $N_{hard}(a,b)$ in (3.4) asymptotically minimize $\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(N_{hard}(a,b))$ (up to the first-order) for each and every post-change hypothesis $(\delta_1, \ldots, \delta_K) \in \Delta$ subject to the false alarm constraint (2.1), as γ in (2.1) goes to ∞. The conclusion also hold if $N_{hard}(a,b)$ is replaced by either the top-r thresholding scheme $N_{top,r}$ in (3.1) or the combined thresholding scheme $N_{comb,r}(a,b)$ in (3.5) when the occurring event affects at most r data streams, i.e., when $(\delta_1, \ldots, \delta_K) \in \Delta$ satisfies $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$.

Proof: This corollary follow at once from Theorems 5.1 and 5.2. In particular, the choice of a_γ in (5.7) follows from (5.5) and the fact that $1 + a + \frac{a^2}{2!} + \cdots + \frac{a^{K-1}}{(K-1)!} \sim \frac{a^{K-1}}{(K-1)!}$ if K is fixed and a goes to ∞. $\square$

It is useful to add several remarks. First, it is worth emphasize the special case $\delta_{\max} = 0$ of Corollary 5.1. While existing statistical methods can provide asymptotically optimal solutions for this scenario, they are generally infeasible in practice. For instance, consider a case that the occurring event is known to affect at most r data streams instantaneously but we do not know the actual subset of affected data streams. To develop asymptotically optimal schemes, the classical generalized likelihood ratio method would require us to search all possible $\binom{K}{1} + \ldots + \binom{K}{r}$ combinations of affected data streams, which can be too huge from the computational viewpoint. On the other hand, the proposed top-r thresholding scheme $N_{top,r}$ in (3.1) is not only asymptotically optimal for this case but also computationally feasible to be implemented for large K over long time period.

One may worry about the term $\delta_{\max}$ in (5.6) at first glance. However, we want to emphasize that this is just some technical details to simplify our arguments. Otherwise we can focus on a subset of $\{1, \ldots, K\} : \mathcal{K} = \{k : \delta_k = O(\log \gamma)\}$, and pretend that we are just monitoring those data streams inside the subset $\mathcal{K}$ without worrying those outside of $\mathcal{K}$. This is because subject to the false alarm constraint γ in (2.1), the detection delays are typically in the order of $\log \gamma$, and the data streams inside $\mathcal{K}$ will raise an alarm at time $\nu + O(\log \gamma)$ before the event affects those data streams outside of $\mathcal{K}$. In other words, the term $\delta_{\max}$ in (5.6) can be easily replaced by $O(\log \gamma)$ (although the term $J(\delta_1, \ldots, \delta_k)$ shall also be defined accordingly, i.e., only consider those δ_k inside $\mathcal{K}$). In addition, in our corollary we make a stronger assumption $\delta_{\max} = o(\log \gamma)$ instead of $O(\log \gamma)$ so that we can take advantage of the well-known asymptotic lower bounds on detection delays subject to the false alarm constraint γ in (2.1). However, this is not essential either, and we conjecture that the proposed thresholding schemes hold their respective asymptotic optimality properties regardless of the value of $\delta_{\max}$.

It is also interesting to see what happens if both K and γ go to ∞. We do not have rigorous mathematical arguments, but our theorems seem to lead to the following heuristic arguments. Note that the right-hand side of (5.5) can be written as $1/P(U_K \geq a)$, where U_K is the sum of K independent exponential random variables with mean 1. By the theory of large deviations, for any constant $b > 1$, we have

$$\lim_{K \to \infty} -\frac{1}{K} \log \mathbf{P}(U_K \geq Kb) = b - 1 - \log(b),$$

see, for example, Durrett [5, Ch. 1.9]. Assume $\log \gamma = K(b-1-\log(b))$, then we can choose the threshold $a = Kb$ to satisfy the false alarm constraint (2.1) if we assume that the lower bound (5.5) still holds when K goes to ∞. By Theorem 5.2, the asymptotic efficiency of our proposed thresholding schemes as compared to the lower bound in Theorem 5.1 will be $b/(b-1-\log b)$, which goes to 1 if b goes to ∞, or equivalently, if $b-1-\log(b) = \log \gamma / K$ goes to ∞. In other words, the asymptotic optimality properties of our proposed thresholding schemes seem to still hold when the number K of data streams and the false alarm constraint γ go to ∞ in such a way that $\log \gamma / K$ also goes to ∞. However, in other scenarios, our proposed schemes no longer achieve the asymptotic lower bound in Theorem 5.1, which may or may not provide a sharp lower bound on the detection delays when K goes to ∞.

6. Discussion. From the methodology viewpoint, our proposed approaches can be easily extended to other reasonable local detection statistics or other more complicated models, as long as the observations are independent among different data streams. As an example, besides the local CUSUM statistics $W_{k,n}$, another widely used local detection statistics is the quasi-Bayesian-type statistic of Shiryaev [25] and Roberts [23]. The Shiryaev-Roberts statistic is defined by

$$R_{k,n} = \sum_{\nu=1}^{n} \prod_{i=\nu}^{n} \frac{g_k(X_{k,i})}{f_k(X_{k,i})},$$

and has a recursive formula: $R_{k,n} = (1 + R_{k,n-1}) \frac{g_k(X_{k,n})}{f_k(X_{k,n})}$ with $R_{k,0} = 0$. If we compare e^{W_n} with R_n, it is easy to see that the maximum over the change-point ν in the CUSUM statistic is replaced by the sum in the Shiryaev-Roberts statistic R_n, and it is well-known that Page's CUSUM and Shiryaev-Roberts procedures, based on $W_{k,n}$ and $\log R_{k,n}$ respectively, have similar performance when monitoring a single local one-dimensional data stream, see, Pollak and Siemgund [21]. Hence, it is natural to ask whether the local CUSUM statistics W_n in our proposed schemes can be replaced by $\log R_{k,n}$, the Shiryaev-Roberts statistic in the logarithm scale. Our preliminary numerical simulation study seems to suggest that the answer is still a "Yes" when monitoring a large number of data streams, as the Shiryev-Roberts like version also has similar performance as the CUSUM version in our simulation study in Section 4.

Likewise, our approaches thresholding schemes can also be easily adjusted for more complicated models, e.g., when the post-change distribution involve unknown parameters. Extensive research has been done in the past decades to deal with more complicated models for one dimensional data stream, see, Pollak and Siemgund [21], Siegmund and Venktraman [27], Lai [9], Lorden and Pollak [12]. As an illustration, assume that the observations at the k-th data stream have a f_k distribution before the change, but may have a g_{k,θ_k} distribution after the change if the k-th data stream is affected by the occurring event, where the value θ_k is unknown and may differ for different k. In this case, several theoretically efficient and computationally simple local schemes have been proposed, and most have the form that raises an alarm if a local detection statistic at time n, say, $W^*_{k,n}$, exceeds some pre-specified constant. To be more specific, let us consider the schemes proposed in Lorden and Pollak [12], in which one first find an estimator $\hat{\theta}_{k,\nu,n-1}$ of θ_k by the maximum likelihood or method of moments methods based on the observations from $X_{k,\nu}, \cdots, X_{k,n-1}$ for each given $1 \le \nu \le n-1$ and then use the estimator $\hat{\theta}_{k,\nu,n-1}$ to construct a local detection statistic $W^*_{k,n}$ that mimics the local CUSUM or Shiryaev-Roberts statistics, e.g.,

$$W^*_{k,n} = \max\left\{0, \max_{1 \le \nu \le n-1} \sum_{i=\nu}^{n} \log \frac{g_{k,\hat{\theta}_{k,\nu,n-1}}(X_{k,i})}{f_k(X_{k,i})}\right\}$$

or

$$W^*_{k,n} = \log \sum_{\nu=1}^{n} \prod_{i=\nu}^{n} \frac{g_{k,\hat{\theta}_{k,\nu,n-1}}(X_{k,i})}{f_k(X_{k,i})}.$$

Then our proposed top-r or hard thresholding rules or both can be applied to these local detection statistic $W^*_{k,n}$ too, thereby providing scalable global schemes in the case when unknown parameters are present in the post-change distributions.

When the local detection statistic is the Shiryaev-Roberts statistics or other complicated local detection statistic, we conjecture that the proposed thresholding schemes still hold certain asymptotic optimality properties. Unfortunately, we have so far been unable to provide a full proof. The main challenge is to establish a lower bound on the average run length to false alarm in parallel to those in (5.5). Another difficulty is how to choose the hard-thresholding b_k's in more complicated models.

Appendix: Proof of Theorems 5.1 and 5.2. This section is devoted to prove our main theorems, Theorems 5.1 and 5.2.

Proof of Theorem 5.1. Intuitively, only those affected data streams provide information to detect the occurring events, and the quickest possible way to detect the occurring event is when the event affects the data streams instantaneously. More rigorously, if we define

$$\delta_k^* = \begin{cases} 0, & \text{if } \delta_k \text{ is finite} \\ \infty, & \text{if } \delta_k = \infty \end{cases}, \tag{6.1}$$

then for any given scheme $T(\gamma)$,

$$\overline{\mathbf{E}}_{\delta_1,\ldots,\delta_K}(T(\gamma)) \geq \inf_{\tau} \overline{\mathbf{E}}_{\delta_1^*,\ldots,\delta_K^*}(\tau),$$

where the infumum is taken over all possible scheme τ satisfying the false alarm constraint (2.1). An alternative and possible better viewpoint is based on a time-shifting argument in which one imagines that at time n one observes the observations $X_{k,n-\delta_k}$ (instead of $X_{k,n}$) when δ_k is finite, and then applies $T(\gamma)$ to the new aligned observations.

Without loss of generality, assume that m out of K data streams are affected by the event, and $\delta_i^* = 0$ for $1 \leq i \leq m$, and $= \infty$ for $m+1 \leq i \leq K$. That is, the first m data streams are affected abruptly and simultaneously by the event at unknown time ν, and other data streams are unaffected. Then it is well-known [16] that the corresponding optimal scheme is the CUSUM procedure

$$T_{CUSUM}^{(m)}(a) = \inf\{n \geq 1 : W_n^{(m)} \geq a\}$$

where the corresponding CUSUM statistic is given by

$$W_n^{(m)} = \max\Big(W_{n-1}^{(m)} + \sum_{k=1}^{m} \log \frac{g_k(X_{k,n})}{f_k(X_{k,n})},\ 0\Big).$$

By (5.4), $J(\delta_1,\ldots,\delta_K) = J(\delta_1^*,\ldots,\delta_K^*) = \sum_{i=1}^{m} I(g_i, f_i)$ and thus the right-hand side of (5.3) is the detection delay of the optimal scheme $T_{CUSUM}^{(m)}(a)$

that provides a lower bound on the detection delays of all other schemes. Therefore, relation (5.3) holds and this completes the proof of Theorem 5.1.

Proof of Theorem 5.2. Let us first focus part (i) on the properties of the hard thresholding scheme $N_{hard}(a, b)$ in (3.4).

To prove (5.5) in part (i), note that each of the proposed thresholding schemes $N_{a,b}$ dominates the "SUM" scheme $T_{\text{sum}}(d = a)$ in (2.5). That is, for any $b \geq 0$, $N_{a,b} \geq T_{\text{sum}}(a)$ and $\mathbf{E}^{(\infty)}(N_{a,b}) \geq \mathbf{E}^{(\infty)}(T_{\text{sum}}(a))$. By Theorem 1 of Mei [14], the "SUM" scheme $T_{\text{sum}}(a)$ satisfies relation (5.5), and so are the proposed thresholding schemes $N_{a,b}$ for all $b \geq 0$.

Now let us prove relation (5.6) in part (ii), and we first focus on the hard thresholding scheme $N_{hard}(a, b)$. It is clear that the worst-case detection delay of $N_{hard}(a, b)$ occurs at the change-point $\nu = 1$, and thus it suffices to show that $\mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b))$ satisfies (5.6). Without loss of generality, we assume that only the first m data steams are affected and no other data streams are affected. To simply our notation below, denote $\delta_{\max} = \max_{1 \leq i \leq m} \delta_i$. Since $N_{hard}(a, b)$ is increasing as a function of $b \geq 0$, it suffices to show that

$$\mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b')) \leq \frac{a}{\sum_{k=1}^{m} I(g_k, f_k)} + O(\sqrt{b'}) + \delta_{\max}, \tag{6.2}$$

as b' and $a - b'$ go to ∞.

To prove (6.2), the essential idea is to compare $N_{hard}(a, b')$ with a new stopping time that starts to monitor the change at time $\delta_{\max}$ and is in the form of the one-sided sequential probability ratio test (SPRT). Define a stopping time

$$\begin{aligned} \tau(a, b') \quad &= \quad \text{first } n \text{ such that } \sum_{i=1}^{n} \sum_{k=1}^{m} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq a \quad \text{and} \\ & \qquad \sum_{i=1}^{n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq \rho_k b' \quad \text{for all } 1 \leq k \leq m, \end{aligned} \tag{6.3}$$

where ρ_k is defined in (3.2), and let $\hat{\tau}_\delta(a, b')$ be the new stopping time that applies $\tau(a, b')$ to the new observations after time $\delta_{\max}$. Clearly, whenever $\hat{\tau}_\delta(a, b')$ stops, our proposed scheme $N_{hard}(a, b')$ also stops (possibly earlier). Thus

$$\begin{aligned} \mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(N_{hard}(a, b')) &\leq \mathbf{E}^{(\nu=1)}_{\delta_1,\ldots,\delta_K}(\hat{\tau}_\delta(a, b')) \\ &= \delta_{\max} - 1 + \mathbf{E}^{(\nu=1)}_{\delta_1^*,\ldots,\delta_K^*}(\tau(a, b')), \end{aligned}$$

where δ_k^* is defined in (6.1). To simplify the notation, denote by $\mathbf{E}^{(1)}$ the expectation when the change occurs at time $\nu = 1$ and the event affects the first m data streams immediately but does not affect the other remaining $K - m$ data streams. So it suffices to show that the stopping time $\tau(a, b')$ in (6.3) satisfies

$$\mathbf{E}^{(1)}(\tau(a, b')) \leq \frac{a}{\sum_{k=1}^m I(g_k, f_k)} + O(\sqrt{b'}). \tag{6.4}$$

To prove this, for $1 \leq k \leq m$, let

$$\begin{aligned} M_k &= \inf\Big\{n \geq 1 : \sum_{i=1}^{n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \geq \rho_k b'\Big\}, \\ \tau_k(M_k) &= \sup\Big\{n \geq 1 : \sum_{i=M_k+1}^{M_k+n} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})} \leq 0\Big\} \\ \hat{M} &= \max_{1 \leq k \leq m} \Big(M_k + \tau_k(M_k) + 1\Big) \\ t(\hat{M}) &= \inf\Big\{n \geq 1 : \sum_{i=\hat{M}+1}^{\hat{M}+n} \Big(\sum_{k=1}^{m} \log \frac{g_k(X_{k,i})}{f_k(X_{k,i})}\Big) \geq a - (\sum_{k=1}^{m} \rho_k) b'\Big\}. \end{aligned}$$

Note that in the definition of $t(\hat{M})$, the threshold $a - (\sum_{k=1}^m \rho_k)b'$ goes to ∞ as $a \to \infty$, since $\sum_{k=1}^m \rho_k \leq \sum_{k=1}^K \rho_k = 1$ and $a - b'$ is assumed to go to ∞. Combining these definitions with those of $\tau(a, b')$ in (6.3) yields that

$$\tau(a, b') \leq \hat{M} + t(\hat{M}) = \max_{1 \leq k \leq m} \Big(M_k + \tau_k(M_k) + 1\Big) + t(\hat{M})$$

$$\leq \sum_{k=1}^{m} \tau_k(M_k) + 1 + t(\hat{M}) + \max_{1 \leq k \leq m} M_k.$$

Hence, to prove (6.4), it suffices to establish the following three relations:

$$\begin{aligned}
&(6.5) \qquad \mathbf{E}^{(1)}\Big(\tau_k(M_k)\Big) &=& \; O(1) \qquad \text{for all } 1 \leq k \leq m; \\
&(6.6) \qquad \mathbf{E}^{(1)}\Big(t(\hat{M})\Big) &\leq& \; \frac{a}{\sum_{k=1}^{m} I(g_k, f_k)} - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + O(1); \\
&(6.7) \; \mathbf{E}^{(1)}\Big(\max_{1 \leq k \leq m} M_k\Big) &\leq& \; \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + O(\sqrt{b'}).
\end{aligned}$$

Relation (6.5) is well-known in renewal theory, e.g., Theorem D in Kiefer and Sacks [7], since $\log\big(g_k(X)/f_k(X)\big)$ has positive mean and finite variance under $\mathbf{E}^{(1)}$ by Assumption (A2). Relation (6.6) also follows at once from the standard renewal theory

$$\begin{aligned}
\mathbf{E}^{(1)}(t(\hat{M})) &= \frac{a - (\sum_{k=1}^{m} \rho_k) b'}{\sum_{k=1}^{m} I(g_k, f_k)} + O(1) \\
&= \frac{a}{\sum_{k=1}^{m} I(g_k, f_k)} - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + O(1),
\end{aligned}$$

as $a - b'$ goes to ∞, where the second equality follows from the definition of ρ_k in (3.2).

The proof of relation (6.7) is a little more complicated, but it can be done along the same line as that in Mei [13]. Specifically, by renewal theory and Assumption (A2), under $\mathbf{P}^{(1)}$,

$$\mathbf{E}^{(1)}(M_k) = \frac{\rho_k b'}{I(g_k, f_k)} + O(1) = \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + O(1)$$

and $\mathrm{Var}^{(1)}(M_k) = O(b')$, as $b \to \infty$, see Siegmund [26, p. 171]. Hence,

$$\begin{aligned}
\mathbf{E}^{(1)}\Big(\max_{1 \leq k \leq m} M_k\Big) &= \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + \mathbf{E}^{(1)} \max_{1 \leq k \leq m} \Big(M_k - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)}\Big) \\
&\leq \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + \sum_{k=1}^{m} \mathbf{E}^{(1)} \Big| M_k - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} \Big|
\end{aligned}$$

$$\leq \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)} + O(\sqrt{b'}),$$

where the last inequality follows from the fact that

$$\begin{aligned}\Big(\mathbf{E}^{(1)}|M_k - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)}|\Big)^2 &\leq \mathbf{E}^{(1)}\Big(M_k - \frac{a}{\sum_{k=1}^{K} I(g_k, f_k)}\Big)^2 \\ &= \mathrm{Var}^{(1)}(M_k) + \Big(\mathbf{E}^{(1)}M_k - \frac{b'}{\sum_{k=1}^{K} I(g_k, f_k)}\Big)^2 \\ &= O(b').\end{aligned}$$

Therefore, relations (6.5)-(6.7) are proved, and hence relation (5.6) holds for the hard-thresholding scheme $N_{hard}(a, b)$ in (3.4).

Now let us provide a sketch of the proof for part (ii) of Theorem 5.2 on the top-r scheme $N_{top,r}(a)$ in (3.1) and the combined thresholding scheme $N_{comb,r}(a, b)$ in (3.5). Since $N_{top,r}(a)$ is a special case of $N_{comb,r}(a, b)$ with $b = 0$, it suffices to prove the theorem for $N_{comb,r}(a, b)$ in (3.5) with $b \geq 0$. The properties of false alarms are straightforward since the centralized "SUM" scheme $T_{\mathrm{sum}}(d)$ again provides the lower bound and thus relation (5.5) also holds for $N_{comb,r}(a, b)$ with $b \geq 0$.

It remains to show that when the occurring event affects at most r data streams, i.e., when $\sum_{k=1}^{K} I\{\delta_k < \infty\} \leq r$, relation (5.6) holds for $N_{comb,r}(a, b)$ with $b \geq 0$. Without loss of generality, assume that the affected data streams are just the first m data streams with $m \leq r$. Recall that $U_{k,n} = W_{k,n} I\{W_{k,n} \geq \rho_k b\} \geq 0$, and we order the non-negative $U_{k,n}$'s as $U_{(1),n} \geq \ldots \geq U_{(K),n}$, and $N_{comb,r}(a, b)$ stops if $\sum_{k=1}^{r} U_{(k),n} \geq a$. When $m \leq r$, we have

$$\sum_{k=1}^{r} U_{(k),n} \geq \sum_{k=1}^{r} U_{k,n} \geq \sum_{k=1}^{m} U_{k,n}.$$

Thus, if at some time n_0 we have $W_{k,n_0} \geq \rho_k b$ and $\sum_{k=1}^{m} W_{k,n_0} \geq a$ for $1 \leq k \leq m$ (i.e., for the first m data streams), then $N_{comb,r}(a, b)$ will also stop at time n_0 and possibly earlier. Hence, whenever $m \leq r$, the stopping time $\tau(a, b')$ in (6.3) also provides an upper bound on the detection delay of $N_{comb,r}(a, b)$. Thus the proposed combined thresholding scheme $N_{comb,r}(a, b)$ in (3.5) satisfies relation (5.6) whenever the occurring event affects at most r data streams. This completes the proof of Theorem 5.2. □

References.

[1] Appadwedula, S., Veeravalli, V. V., and Jones, D. (2005). Energy-efficient detection in sensor networks. *IEEE J. Sel. Areas Commun.*, *23*, 693–702.

[2] Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Applications.* Englewood Cliffs, Prentice-Hall. MR1210954

[3] Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica*, *15*, 257–325. MR2269743

[4] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[5] Durrett, R. (1996). *Probability: Theory and Examples.* Second edition. Duxbury Press, Belmont, CA. MR1609153

[6] Fan, J. and Lin, S. K. (1998). Test of significance when data are curves. *Journal of American Statistical Association*, **93**, 1007–1021.

[7] Kiefer, J. and Sacks, J. (1963). Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34** 705–750. MR0150907

[8] Kulldorff, M. (2001). Prospective Time-Periodic Geographic Disease Surveillance Using a Scan Statistic, *Journal of Royal Statistical Society, Series A* **164** 61–72.

[9] Lai, T. L. (1995). Sequential change-point detection in quality control and dynamical systems (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 613–658. MR 1672051

[10] Lai, T. L. (2001). Sequential analysis: some classical problems and new challenges. *Statist. Sinica* **11** 303–408. MR1844531

[11] Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42** 1897–1908. MR0309251

[12] Lorden, G. and Pollak M. (2005). Nonanticipating estimation applied to sequential analysis and changepoint detection. *Ann. Statist.* **33** 1422–1454. MR2195641

[13] Mei, Y. (2005). Information bounds and quickest change detection in decentralized decision systems. *IEEE Trans. Inform. Theory* **51** 2669–2681. MR2246385

[14] Mei, Y. (2010). Efficient scalable schemes for moniotoring a large number of data streams. *Biometrika.* **97** 419–433.

[15] Montgomery, D. C. (1991). *Introduction to Statistical Quality Control* (2nd edition). Wiley, New York.

[16] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14** 1379–1387. MR0868306

[17] Neyman, J. (1937). Smooth test for goodness-of-fit. *Skand. Aktuarietidskr.* **20** 149–199.

[18] Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850

[19] Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13** 206–227. MR0773162

[20] Pollak, M. (1987). Average run lengths of an optimal method of deteting a change indistribution. *Ann. Statist.* **15** 749–779. MR0888438

[21] Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Ann. Statist.* **19** 394–416. MR1091859

[22] Rago, C., Willett, P., and Bar-Shalom, Y. (1996). Censoring sensors: A low-communication-rate scheme for distributed detection. *IEEE Trans. Aerosp. Electon. Syst.*, *32*, 554–568.

[23] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics* **8** 411–430. MR0196887

[24] Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product.* D Van Norstrand, New York. Preprinted by ASQC Quality Press, Wisconsin, 1980.

[25] Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.* **8** 22–46.

[26] Siegmund, D. (1985): *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York. MR0799155

[27] Siegmund, D. and Venktraman, E. S. (1991). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.* **23** 255–271. MR1331667

[28] Tartakovsky, A. G. and Veeravalli, V. V. (2004). Change-point Detection in

Multichannel and Distributed Systems. *Applied Sequential Methodologies*, 339–370, Statist. Textbooks Monogr., 173, Dekker, New York. MR2159163

[29] Tartakovskya, A. G., Rozovskiia, B. L., Blazeka, R. B. and Kim, H. (2006). Detection of intrusions in information systems by sequential change-point methods (with discussions). *Statistical Methodology* **3** 252–340. MR2240956

[30] Tay, W.-P., Tsitsiklis, J. N. and Win, M. Z. (2007). Asymptotic performance of a censoring sensor network. *IEEE Trans. Inform. Theory* **53** 4191–4209.

[31] Veeravalli, V. V. (2001). Decentralized quickest change detection. *IEEE Trans. Inform. Theory* **47** 1657–1665.

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA
E-mail: ymei@isye.gatech.edu

4.2 Manuscript B: A Generalization of Kullback-Leibler's Inequality and Its Applications to Quantization Effects on Detection.

A Generalization of Kullback-Leibler's Inequality and Its Applications to Quantization Effects on Detection

Yan Wang, Yajun Mei

Abstract

It is well known that quantization cannot increase the Kullback-Leibler divergence which can be thought of as the first moment of the log-likelihood ratios. In this paper, this result is extended to the case of second and higher moments of the log-likelihood ratios. It is shown that quantization may result in an increase of the second or higher moments of the log-likelihood ratio, but such an increase is at most by a universal constant that only depends on the value of the moment. Such a constant is $\mathbf{2/e}$ for the second moment. The result is further applied to decentralized quickest detection problems to provide a simpler sufficient condition for the asymptotic optimality theory.

I. Introduction

Let X be a random variable taking value in some probability space $(\Omega, \mathcal{F}, P)$. In some applications, the X itself is unobservable and what is actually observed is another variable Y that is a quantization of X, or more generally, a function of X, say $Y = \phi(X)$. Thus one has to utilize the Y to develop statistical procedures to make decisions. In order to guarantee the success, one may need to verify that the distribution of Y satisfies some necessary requirements. However, it can be analytical challenging or untractable to verify the assumptions for Y directly, even if the distributions of the unobservable X are known to belong to some simple family of distributions. For instance, this may happen when one only has very limited knowledge about the function ϕ, e.g., ϕ belongs to some infinite-dimensional functional space. To overcome such a difficulty, it is natural to investigate whether certain properties of the X will make sure that the requirements of Y are satisfied, see Le Cam and Yang [5] for more detailed discussions and some concrete examples.

Our research is motivated by the decentralized sequential detection where one wants to find an appropriate quantization function ϕ so as to make the best possible decision subject to the constraint that the observed variables Y's belong to a finite alphabet. This requires us to investigate the properties of the moments of log-likelihood ratio statistic under quantization or mapping. Suppose that $H_0 : P = P_0$ and $H_1 : P = P_1$ are two simple hypotheses regarding the distribution P of the X, and the X has a density $f_i(x)$ under H_i (or P_i) with respect to some common underlying probability measure. In information theory and statistics, the log-likelihood ratio of X, or the logarithm of Radon-Nikodym derivative of P_1 with respect to P_0, is defined by

$$Z = \log \frac{dP_1}{dP_0}(X) = \log \frac{f_1(X)}{f_0(X)},$$

and the widely used Kullback-Leibler divergence of the X, denoted by $I(f_1, f_0)$, is just the first moment of Z under P_1, i.e., $I(f_1, f_0) = E_1\{Z\}$. Likewise, for the $Y = \phi(X)$, denote by P_i^{ϕ} and $f_i(y;\phi)$ its probability measure and its probability mass or density function $f_i(y;\phi)$ under H_i. Then the log-likelihood ratio of $Y = \phi(X)$ is given by

$$Z_\phi = \log \frac{dP_1^{\phi}}{dP_0^{\phi}}(Y) = \log \frac{f_1(Y;\phi)}{f_0(Y;\phi)}.$$

and the Kullback-Leibler divergence of the Y is $I_\phi(f_1, f_0) = E_1\left\{Z_\phi\right\}$. An important property is that the Kullback-Leibler divergence cannot increase under a mapping, that is,

$$I_\phi(f_1, f_0) \leq I(f_1, f_0) \tag{1}$$

with equality if and only if $Y = \phi(X)$ is a sufficient statistics of X, see Theorem 4.1 of Kullback and Leibler [3]. This is consistent with our intuition that $Y = \phi(X)$ is generally less informative than the X itself. Note that the inequality (1), which will be referred as *Kullback-Leibler's inequality* below, deals with the first moment (or mean) of the log-likelihood ratios.

The main goal of the present paper is to extend Kullback-Leibler's inequality (1) to deal with the second or other higher order moments of log-likelihood ratios. Section II extends the Kullback-Leibler's inequality (1) to second moments, and Section III further extends it to other general higher-order moments. In Section IV, our results are applied to decentralized sequential detection to provide much simplified sufficient conditions for asymptotic optimality theories. Section V gives concluding remarks.

II. Second-order Moments

For the X and Y, define their respective second moments of log-likelihood ratios as

$$V(f_1, f_0) = E_1\left\{Z^2\right\} = E_1\left\{\left(\log \frac{f_1(X)}{f_0(X)}\right)^2\right\}$$

and

$$V_\phi(f_1, f_0) = E_1\left\{Z_\phi^2\right\} = E_1\left\{\left(\log \frac{f_1(Y;\phi)}{f_0(Y;\phi)}\right)^2\right\}.$$

Our main result is as follows.

Theorem 1. *For any measurable function ϕ, we have*

$$V_\phi(f_1, f_0) \leq V(f_1, f_0) + \frac{2}{e}. \tag{2}$$

Proof: Let $L = e^{Z} = f_1(X)/f_0(X)$ and $L_\phi = e^{Z_\phi} = f_1(Y;\phi)/f_0(Y;\phi)$ be the likelihood ratios. Recall that in the proof of the Kullback-Leibler's inequality (1), a key idea is to use the fact that the function $H(t) = -\log t$ is convex, so that we can apply Jensen's inequality to $Z_\phi = \log L_\phi = -H(L_\phi)$. However this approach fails for the second moment case since the function $H_2(t) = (-\log t)^2 = (\log t)^2$ is no longer convex (or concave). Fortunately, this idea can be salvaged by finding a convex function that dominates $H_2(t)$. Specifically, note that the function $H_2(t) = (\log t)^2$ is convex on $t \le e$ but is concave on $t \ge e$, and thus we can consider a new function

$$\tilde{H}_2(t) = \begin{cases} (\log t)^2, & \text{if } 0 < t \le e \\ \frac{2}{e}t - 1 & \text{if } t > e \end{cases}. \tag{3}$$

Then it is easy to see that $\tilde{H}_2(t)$ is a continuous convex function of t when $t \ge 0$. Moreover, $\tilde{H}_2(t) \ge H_2(t) = (\log t)^2$. In fact, $\tilde{H}_2(t)$ equals to $H_2(t)$ when $t \le e$ and becomes linear when $t \ge e$, see Fig. 1.

To prove our theorem, let $E_1\{\cdot|Y\}$ denote the conditional expectation with respect to a given value of the observed data $Y = \phi(X) \in \Theta$, then

$$E_1\left\{ L^{-1} \middle| Y\right\} = E_1\left\{ \frac{f_0(X)}{f_1(X)} \middle| Y\right\} = \frac{f_0(Y;\phi)}{f_1(Y;\phi)} = L_\phi^{-1}.$$

Since $H_2(t) = (\log t)^2 = H_2(t^{-1})$ for $t > 0$, by the definition of $V_\phi(f_1, f_0)$, we have

$$\begin{aligned} V_\phi(f_1, f_0) &= E_1\{(\log(L_\phi))^2\} = E_1\left\{H_2(L_\phi^{-1})\right\} \\ &\le E_1\left\{\tilde{H}_2(L_\phi^{-1})\right\} \\ &= E_1\left\{\tilde{H}_2\left(E_1\left\{ L^{-1} \middle| Y\right\}\right)\right\} \\ &\le E_1\left\{E_1\left\{\tilde{H}_2\left(L^{-1}\right) \middle| Y\right\}\right\} \\ &= E_1\left\{\tilde{H}_2\left(L^{-1}\right)\right\}. \end{aligned}$$

Now by the definition of $\tilde{H}_2(t)$ on equation (3),

$$\begin{aligned} E_1\left\{\tilde{H}_2\left(L^{-1}\right)\right\} &= E_1\left\{\tilde{H}_2\left(L^{-1}\right) I\left\{L^{-1} \le e\right\}\right\} + E_1\left\{\tilde{H}_2\left(L\right) I\left\{L^{-1} > e\right\}\right\} \\ &= E_1\left\{H_2\left(L^{-1}\right) I\left\{L^{-1} \le e\right\}\right\} + E_1\left\{\left(\frac{2}{e}L^{-1} - 1\right) I\left\{L^{-1} > e\right\}\right\} \\ &\le E_1\left\{H_2\left(L^{-1}\right)\right\} + \frac{2}{e}E_1\left\{L^{-1}\right\} - P_1\left\{L^{-1} > e\right\} \\ &\le V(f_1, f_0) + \frac{2}{e} \end{aligned}$$

where the last inequality follows from the facts that $V(f_1, f_0) = E_1\{H_2(L^{-1})\}$, $E_1\{L^{-1}\} = \int (f_0(x)/f_1(x)) f_1(x) dx = 1$ and $P_1(L^{-1} > e) \ge 0$. Combining the above inequalities yields (2), completing the proof of the theorem. ■

It is useful to provide some comments to better understand our theorems. First, the discrete version of the Kullback-Leibler's inequality (1) is the well-known log-sum inequality: for nonnegative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, denote the sum of all a_i's by a and the sum of all b_i's by b, and then we have

$$a \log \frac{a}{b} \le \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}$$

with equality if and only if a_i/b_i are constant. Meanwhile, the discrete version of our main result (2) becomes that

$$a\Big(\log \frac{a}{b}\Big)^2 \le \sum_{i=1}^{n} a_i \Big(\log \frac{a_i}{b_i}\Big)^2 + \frac{2}{e} b,$$

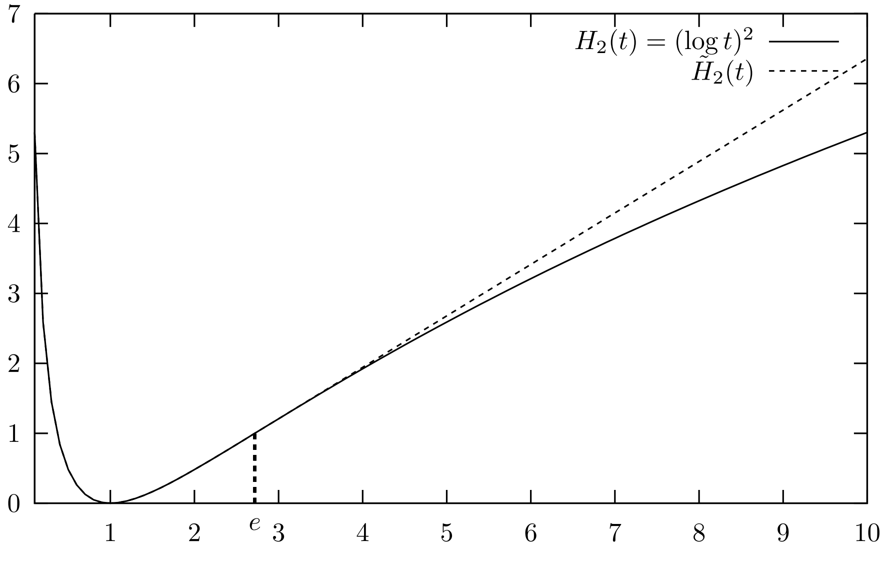

Fig. 1: Dominating Function $\tilde{H}_2(t)$

where the extra term on the right side is $2b/e$ instead of $2/e$ since we do not put any normalization conditions on a or b.

Second, a comparison of (1) and (2) shows that we have an extra constant term $2/e$ for the second moment case, and thus it is natural to ask whether or not the term can be eliminated, i.e., whether it is always true that $V_\phi(f_1, f_0) \leq V(f_1, f_0)$. The following counterexample provides a negative answer. Suppose that the X takes three distinct values 0, 1, 2 with probabilities 29/36, 1/9, 1/12 under P_1 and equal probabilities $1/3$ under P_0. Let ϕ be a function with a binary range $\{0, 1\}$ such that $\phi(0) = 0$, $\phi(1) = \phi(2) = 1$. Then it is easy to verify that $V(f_1, f_0) = 0.9215 \leq V_\phi(f_1, f_0) = 0.9224$. More generally, other counterexamples can be easily found by choosing two distributions P_1 and P_0 of X, both of which are supported on $n+1$ ($n \geq 2$) points $x_0, \ldots, x_n$ such that the likelihood ratio $L_0 = f_0(x_0)/f_1(x_0) < e$ and $L_i = f_0(x_i)/f_1(x_i) > e$ for $i = 1, \ldots, n$ with $L_1, \ldots, L_n$ being n distinct values. Then if we consider a function ϕ that maps all $x_1, \ldots, x_n$ to a single point y_1 but maps x_0 to another point y_0, then $V(f_1, f_0) \leq V_\phi(f_1, f_0)$. To see this, note that $H_2(t) = (\log t)^2$ is strictly concave on $t \geq e$, so

$$\begin{aligned}
& \sum_{i=1}^{n} f_1(x_i) H_2\left(\frac{f_0(x_i)}{f_1(x_i)}\right) \\
< \quad & (1 - f_1(x_0)) H_2\left(\sum_{i=1}^{n} \frac{f_1(x_i)}{1 - f_1(x_0)} \frac{f_0(x_i)}{f_1(x_i)}\right) \\
= \quad & f_1(y_1; \phi) H_2\left(\frac{f_0(y_1; \phi)}{f_1(y_1; \phi)}\right),
\end{aligned}$$

and

$$\begin{aligned} V(f_1, f_0) &= \sum_{i=0}^{n} f_1(x_i) H_2\left(\frac{f_0(x_i)}{f_1(x_i)}\right) \\ &< f_1(x_0) H_2\left(\frac{f_0(x_0)}{f_1(x_0)}\right) + f_1(y_1;\phi) H_2\left(\frac{f_0(y_1;\phi)}{f_1(y_1;\phi)}\right) \\ &= V_\phi(f_1, f_0). \end{aligned}$$

In other words, unlike the case of Kullback-Leibler's inequality, a map indeed can inflate the second moment of the log-likelihood ratio. Fortunately, our theorem shows that such an inflation is at most $2/e$.

III. General Higher-order Moments

The technique we developed in proving Theorem 1 can be useful to deal with higher-order moments of the log-likelihood ratios. To be specific, for a positive integer $j = 1, 2, \ldots,$ define

$$W_j(f_1, f_0) = E_1\left\{(Z)^j\right\} = E_1\left\{\left(\log \frac{f_1(X)}{f_0(X)}\right)^j\right\} \tag{4}$$

and

$$W_{\phi,j}(f_1, f_0) = E_1\left\{(Z_\phi)^j\right\} = E_1\left\{\left(\log \frac{f_1(Y;\phi)}{f_0(Y;\phi)}\right)^j\right\}. \tag{5}$$

It turns out that we need to consider two different cases, depending on whether j is even or odd. For the purpose of our theorem, let us define two sequence of constants. For any integer $j \geq 1$, define

$$C_j = \frac{j(j-1)^{j-1}}{e^{j-1}}$$

and when j is odd, further define C_j^* to be the only real number $x \geq 0$ that satisfies the equation

$$x = (j-1)^{j-1} - C_j \exp(-x^{1/j}). \tag{6}$$

By convention we set $0^0 = 1$, and thus $C_1 = 1$ and $C_1^* = 0$.

The following theorem involves higher-order moments of the log-likelihood ratios, and includes the Kullback-Leibler's inequality (1) and relation (2) for second-order moment as special cases.

Theorem 2. *For any measurable function ϕ and any integer $j \geq 1$, we have*

$$W_{\phi,j}(f_1, f_0) \leq W_j(f_1, f_0) + B, \tag{7}$$

where the constant $B = C_j$ if j is even and $B = C_j^*$ if j is odd. Moreover, $W_{\phi,j}(f_1, f_0)$ and $W_j(f_1, f_0)$ have a lower bound 0 when j is even, and have a lower bound $-j(j-1)^{j-1}/e^{j-1} - (j-1)^j$ when j is odd.

We will prove Theorem 2 in two separate cases, depending on whether j is even or odd. Let us begin with the case when j is even, and we will prove a more general result on the α-moments of the absolute values of the log-likelihood ratios Z and Z_ϕ for any real number $\alpha \geq 1$. Specifically, define

$$\tilde{W}_\alpha(f_1, f_0) = E_1\{|Z|^\alpha\} = E_1\left\{\left|\log \frac{f_1(X)}{f_0(X)}\right|^\alpha\right\}$$

and

$$\tilde{W}_{\phi,\alpha}(f_1,f_0) = E_1\left\{|Z_\phi|^\alpha\right\} = E_1\left\{\left|\log\frac{dP_1^\phi}{dP_0^\phi}(Y)\right|^\alpha\right\} = E_1\left\{\left|\log\frac{f_1(Y;\phi)}{f_0(Y;\phi)}\right|^\alpha\right\}.$$

Lemma 1. *For any $\alpha \geq 1$,*

$$\tilde{W}_{\phi,\alpha}(f_1,f_0) \leq \tilde{W}_\alpha(f_1,f_0) + C_\alpha, \tag{8}$$

where the constant $C_\alpha = \frac{\alpha(\alpha-1)^{\alpha-1}}{e^{\alpha-1}} > 0$ and $C_1 = 1$ by convention that $0^0 = 1$.

Proof: While the function $H_\alpha(t) = |\log t|^\alpha$ is not convex, it can be shown that it is dominated by the following convex function

$$\tilde{H}_\alpha(t) = \begin{cases} H_\alpha(t), & \text{if } 0 < t \leq t_\alpha, \\ C_\alpha t - d_\alpha & \text{if } t > t_\alpha, \end{cases}$$

where $t_\alpha = e^{\alpha-1}$, $d_\alpha = (\alpha-1)^{\alpha-1} \geq 0$, and $C_\alpha = \frac{\alpha(\alpha-1)^{\alpha-1}}{e^{\alpha-1}} > 0$. The remaining proof is identical to those of Theorem 1 and thus omitted. ■

As in Theorem 1, it is generally not true that $\tilde{W}_{\phi,\alpha}(f_1,f_0) \leq \tilde{W}_\alpha(f_1,f_0)$, and the counterexamples can be easily found by exploring the fact that for any $\alpha \geq 1$, the function $H_\alpha(t)$ is always strictly concave when $t \geq t_\alpha$. In other words, the counterexamples can be constructed by picking $n+1$ $(n \geq 2)$ points $x_0, \ldots, x_n$ and two distributions P_1 and P_0 such that $L_0 = f_0(x_0)/f_1(x_0) < t_\alpha$ while $L_i = f_0(x_i)/f_1(x_i) > t_\alpha$ are n distinct values for $i = 1, \ldots, n$, and then proceeding as in the case of $\alpha = 2$.

It is also interesting to compare the Kullback-Leibler's inequality (1) with the case $\alpha = 1$ in Lemma 1: we have $E_1\{Z_\phi\} \leq E_1\{Z\}$ and $E_1|Z_\phi| \leq E_1|Z| + 1$. In other words, while the first moment of the log-likelihood ratio always decrease after a mapping, the first moment of its absolute value can indeed increase although such an increase is at most 1. This is because the function $-\log t$ is convex on $t > 0$ but the function $|\log t|$ is not convex.

Now let us prove Theorem 2 when $j \geq 1$ is odd. Fix the odd integer $j \geq 1$, and the key is to find a convex function that dominates $H(t) = (-\log t)^j$. By taking derivatives, it is easy to see that $H(t) = (-\log t)^j$ is convex on $0 < t \leq 1$ or $t \geq e^{j-1}$ but is concave when $1 \leq t \leq e^{j-1}$. Thus, if we let $t_0 = e^{j-1}$, then $H(t) \leq H(t_0) + H'(t_0)(t-t_0) = -C_j t + d_j$ when $1 \leq t \leq t_0$, where $C_j = \frac{j(j-1)^{j-1}}{e^{j-1}} > 0$ and $d_j = (j-1)^{j-1} \geq 0$. A simple calculation shows that the line $y = -C_j t + d_j$ intersects the curve $y = H(t)$ at two points: one of them is $t = t_0 = e^{j-1} \geq 1$ and the other one is in the interval $(0, 1]$ and denoted by $t^* \leq 1$. Therefore, we can construct a function $\tilde{H}(t)$ which dominates $H(t)$:

$$\tilde{H}(t) = \begin{cases} H(t) = (-\log t)^j & \text{if } 0 < t \leq t^*(\leq 1) \\ -C_j t + d_j & \text{if } t^* \leq t < t_0 = e^{j-1} \\ H(t) = (-\log t)^j & \text{if } t \geq t_0(\geq 1) \end{cases}.$$

More importantly, the function $\tilde{H}(t)$ is a convex function on $t > 0$.

Recall the definition of C_j^* in (6) and we claim that $\tilde{H}(t) - H(t) \leq C_j^*$. To prove this claim, first note that $C_j^* = H(t^*) \geq 0$ and it suffices to prove the claim when $t^* \leq t < t_0$ (and thus $\tilde{H}(t)$ is a linear function with negative slope). The proof needs to consider two scenarios, depending on whether $t \leq 1$ or ≥ 1. If $t^* \leq t \leq 1$, then $\tilde{H}(t) \leq \tilde{H}(t^*) = C_j^*$ while $H(t) \geq 0$, and thus the claim holds. Meanwhile, if $1 \leq t < t_0$, then by taking derivatives, it is easy to show that $\tilde{H}(t) - H(t)$ is a decreasing function, and thus

$$\tilde{H}(t) - H(t) \leq \tilde{H}(1) - H(1) = \tilde{H}(1) \leq \tilde{H}(t^*) = C_j^*.$$

Therefore, our claim holds and $0 \leq \tilde{H}(t) - H(t) \leq C_j^*$ for any $t > 0$.

Relation (7) for an odd integer $j \geq 1$ can then be easily proved along the same line as in Theorem 1. It remains to show that for an odd integer $j \geq 1$, $W_j(f_1, f_0)$ in (4) and $W_{\phi,j}(f_1, f_0)$ in (5) are bounded below, since the random variables Z^j or Z_ϕ^j may take positive and negative values. For any random variable X, let $X_+ = \max\{X, 0\}$ be the positive part of X and let $X_- = -\min\{X, 0\}$ be the negative part of X. Then $X = X_+ - X_-$, and it is evident that $X \geq -X_-$. The following lemma completes the proof of Theorem 2.

Lemma 2. *When $j \geq 1$ is an odd integer,*

$$E_1\left\{\left(Z^j\right)_-\right\} \leq j(j-1)^{j-1}/e^{j-1} + (j-1)^j$$

where $0^0 = 1$ by convention.

Proof: Fix the odd integer $j \geq 1$, consider the function

$$\psi(t) = -\min\{0, (-\log t)^j\} = \max\{0, (\log t)^j\}. \tag{9}$$

By taking derivatives, it is easy to see that as a non-decreasing function, $\psi(t)$ is concave on $t \geq t_0$, where $t_0 = e^{j-1}$. Thus

$$\psi(t) \leq \begin{cases} \psi(t_0) & \text{if } t \leq t_0 \\ \psi(t_0) + \psi'(t_0)(t - t_0) & \text{if } t \geq t_0 \end{cases},$$

or equivalently,

$$\psi(t) \leq (j-1)^j I\{t \leq t_0\} + (C_j t - d_j) I\{t > t_0\}$$

where $C_j = \frac{j(j-1)^{j-1}}{e^{j-1}} > 0$ and $d_j = (j-1)^{j-1} \geq 0$. Recall that $L = e^Z = f_1(X)/f_0(X)$ is the likelihood ratio, and thus

$$\begin{aligned} E_1\left\{\left(Z^j\right)_-\right\} &= E_1\left\{\psi\left(L^{-1}\right)\right\} \\ &\leq (j-1)^j P_1\left\{L^{-1} \leq t_0\right\} + C_j E_1\left\{L^{-1} I\{L > t_0\}\right\} - d_j P_1\left\{L^{-1} > t_0\right\} \\ &\leq (j-1)^j + C_j E_1\left\{L^{-1}\right\} \\ &= (j-1)^j + C_j, \end{aligned}$$

completing the proof of the lemma. ■

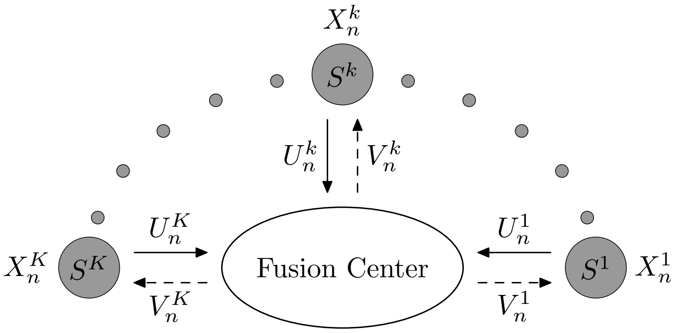

Fig. 2: A Decentralized Sensor Network

IV. Decentralized Sequential Detection

The problem that motivated us to write the present paper arises from decentralized sequential detection, see, Veeravalli, Basar and Poor [18], and Veeravalli [14], [16]. Fig. 2 depicts a typical configuration of a decentralized network system that consists of K geographically deployed local sensors $S^1, \ldots, S^K$ and a fusion center. Each local sensor S^k observes a raw data X_n^k at time step $n = 1, 2, \ldots$, whereas the fusion center makes a final decision when stopping taking observations. Due to constraints on communication bandwidths or requirements of reliability, the local sensors are required to compress the raw data to quantized sensor messages U_n^k's, which all belong to finite alphabets, say $\{0, 1, \ldots, l^k - 1\}$ respectively, and then send the quantized messages to the fusion center. In other words, the fusion center has no direct access to the raw observations and has to make its decisions based on the quantized sensor messages.

There are many possible topologies for the decentralized network system, and one widely used scenario is the system with *limited local memory and full feedback*, or *Case E* in Veeravalli, Basar and Poor [18]. Mathematically, in this scenario, at time n the quantized sensor message from the local sensor S^k to the fusion center is

$$U_n^k \quad = \quad \phi_n^k(X_n^k; \mathcal{F}_{n-1}), \tag{10}$$

where $\mathcal{F}_{n-1} = \{U_{[1,n-1]}^{k=1,\ldots,K}\}$ (here $U_{[1,n]}^k = \{U_1^k, \ldots, U_n^k\}$) denotes all past sensor messages at the fusion center.

In the simplest version of decentralized quickest change detection problems, we assume that an event occurs to the network system at some unknown time ν, and changes the probability measure of the raw data X_n^k from one given probability measure P_0 (with density f_0^k) to another given probability measure P_1 (with density f_1^k). Furthermore, we assume that the observations are independent over time and from sensor to sensor. The objective is how to jointly optimize the policies at the local sensors and fusion center levels so as to detect the change as soon as possible subject to a constraint on the false alarm rate.

A crucial challenge in decentralized quickest change detection is which kind of local quantizers should be used at each local sensor. On the one hand, this is easy if one further assumes that each local sensor uses a stationary local quantizer, as the corresponding problem reduces to the classical centralized case and various well-developed optimal or asymptotic optimal theories are applicable, see for example Lorden [6], Moustakides [9], Page [10], Pollak [11], Shiryayev [12] and [13], etc. In fact, it is not difficult to see that the optimal stationary quantizer ϕ^* for any local sensor S^k is the one that maximizes the local Kullback-Leibler divergence $I_\phi(f_1^k, f_0^k)$, and it can be shown that such an optimal quantizer ϕ^* is a Monotone Likelihood Ratio Quantizer (MLRQ), see, for example, Tsitsiklis [15], Crow and Schwartz [1], Tartakovsky and Veeravalli [14].

On the other hand, the scenario becomes more complicated if the local quantizers are allowed to be non-stationary. By comparing with Bayes procedures, Veeravalli [17] conjectures that the schemes based on the optimal stationary MLRQ ϕ^* are asymptotically optimal regardless whether the quantizers are stationary or not. While this conjecture sounds reasonable as maximizing the Kullback-Leibler divergence seems to be natural to construct optimal local quantizers, it is very challenging to prove or disprove it, partly because of the regularity conditions of the quantized observations. For example, a sequence of non-stationary quantizers may outperform that of stationary quantizers when the second order moments of the log-likelihood ratios of non-stationary quantizers can go to infinity.

Some sufficient conditions under which this conjecture holds are available in the literature. By Lai [4], this conjecture is true under the following sufficient conditions:

$$\lim_{n\to\infty} \sup_{\nu\geq 1} \operatorname{ess\,sup} P^{(\nu)} \left\{ \max_{t\leq n} \sum_{i=\nu}^{\nu+t} \sum_{k=1}^{K} Z_{i,\phi}^k \geq I_{tot}(1+\delta)n \,\middle|\, U_1, \ldots, U_{\nu-1} \right\} = 0. \tag{11}$$

where $P^{(\nu)}$ is the probability measure when the change occurs at time ν, $Z_{i,\phi}^k$ is the likelihood ratio for the quantized data U_i^k, i.e.,

$$Z_{i,\phi}^k = \log \frac{f_1^k(U_i^k; \phi_i^k)}{f_0^k(U_i^k; \phi_i^k)}$$

and $I_{tot} = \sum_{k=1}^K I_{\max}^k$ with $I_{\max}^k = \sup_\phi I(f_1^k, f_0^k; \phi)$. Here $f_m^k(u; \phi_i^k)$ is probability mass function, i.e.,

$$f_m^k(u; \phi_i^k) = P_m^k \left\{ \phi_i^k(X_i^k) = u \right\}, \quad m = 0, 1.$$

Unfortunately, condition (11) involves all possible non-stationary quantizers, and it is impossible to verify it directly. By using Kolmogorov's inequality for martingales, Mei [7] provides a stronger sufficient condition, and shows that the conjecture holds if there is a uniform bound on the second moments of the log-likelihood ratios of quantized observations. Specifically, Mei [7] showed that condition (11) holds if for all $k = 1, \ldots, K$,

$$\sup_\phi V_\phi(f_1^k, f_0^k) < \infty. \tag{12}$$

Moreover, condition (12) holds when the quantized messages belong to binary sensor messages with $l = 2$ and when f_0 and f_1 belong to the same one-parameter exponential family satisfying certain restrictions, see Theorem 2 of [7]. However, it is still an open problem whether condition (12) holds in general or not, as the quantizers can have arbitrary forms and belong to the infinite dimensional functional space.

Our main theorem allows us to tackle more general scenarios. Specifically, by Theorem 1, if for all $k = 1, \ldots, K$,

$$V(f_1^k, f_0^k) = \int \left(\log \frac{f_1^k(x)}{f_0^k(x)} \right)^2 f_1^k(x) dx < \infty, \tag{13}$$

then condition (12) holds and so does (11). Note that condition (13) only deals with the densities f_i^k of raw observations and does not involve the stationary or non-stationary quantizers. Moreover, it is a standard assumption in the statistical literature as a regularity condition for the raw density functions. Therefore, condition (13) provides a simple and reasonable sufficient condition under which the long-standing conjecture of asymptotic optimality of the schemes with the optimal stationary MLRQ ϕ^* is true regardless whether the quantizers are stationary or not.

Similarly, our results can also be applied to the problem of decentralized sequential hypotheses testing, see Veeravalli, Basar and Poor [18]. This problem is similar to the above-mentioned decentralized quickest change detection problem except that the distributions of the raw data do not change over time. In other words, the raw observations $\{X_n^k\}$ form i.i.d. sequences over time n and are independent among different sensors. We still have two simple hypotheses H_0 and H_1 regarding the distributions of X_n^k's, but the objective is to use as few samples as possible to correctly decide which of these two simple hypotheses is true. An optimal sequential test is one that balances the tradeoff between the average sample size under each hypothesis and the probabilities of making Type I and II errors, see Veeravalli, Basar and Poor [18], Veeravalli [16] and Mei [8] for more details.

Unlike the quickest change detection problem, non-stationary quantizers are generally necessary in order to develop asymptotically optimal decentralized sequential tests. This is because each local sensor will have two kinds of optimal stationary quantizers: one maximizes $I_\phi(f_0^k, f_1^k)$ and the other maximizes $I_\phi(f_1^k, f_0^k)$, due to the asymmetric properties of the Kullback-Leibler divergences. Denote them by ϕ_0^k and ϕ_1^k respectively. To develop a simple but asymptotically optimal decentralized sequential tests, Mei [8] introduces the concept of "tandem quantizers" where the test procedure is divided into two stages (also see Section IV of Kiefer and Sacks [2] for a closely related experimental design problem). In the first stage, any reasonable stationary quantizer is used and the network system makes a preliminary decision about which of two hypothesis is likely to be true. Then at the second stage, each local sensor switches to one of two optimal stationary quantizers ϕ_0^k or ϕ_1^k, based on the preliminary decision.

It was shown in Mei [8] that under the condition (12) together with the symmetric condition with f_0 and f_1 exchanged, the decentralized sequential tests with the tandem quantizers are asymptotic optimal among all decentralized sequential tests with or without stationary quantizers. By our Theorem 1, sufficient conditions for the asymptotic optimality of tests with tandem quantizers can be reduced to (13) and the symmetric condition with f_0 and f_1 exchanged.

V. Conclusion

In this paper, we extend the Kullback-Leibler's inequality (1) to deal with the quantization or mapping effects on second or other higher moments of the log-likelihood ratios. Our main results, Theorems 1 and 2, show that a quantization can increase these quantities by at most a universal constant that does not depend on the distributions of the raw observations or the forms of the quantizers. The results are then used to provide a simple but useful sufficient condition for asymptotic optimality theories in decentralized sequential detection.

References

[1] R. W. Crow, S. C. Schwartz, “Quickest detection for sequential decentralized decision systems,” *IEEE Trans. Aerosp. Electron. Syst.,* vol. 32, pp. 267-283, Jan. 1996.

[2] J. Kiefer, J. Sacks, “Asymptotically optimal sequential inference and design,” *Ann. Math. Statist.*, vol. 34, pp. 705-750, 1963.

[3] S. Kullback, R. A. Leibler, “On information and sufficiency”, *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79-86, 1951.

[4] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” *IEEE Trans. Inf. Theory,* vol. 44, pp. 2917-2929, Nov. 1998.

[5] L. Le Cam and G. Yang, “On the preservation of local asymptotic normality under information loss,” *Ann. Statist.,* vol. 16, pp. 483-520, 1988.

[6] G. Lorden, “Procedures for reacting to a change in distribution,” *Ann. Math. Statist.,* vol. 42, pp. 1897-1908, 1971.

[7] Y. Mei, “Information bounds and quickest change detection in decentralized decision systems,” *IEEE Trans. Inf. Theory,* vol. 51, pp. 2669-2681, Jul. 2005.

[8] Y. Mei, “Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks” *IEEE Trans. Inf. Theory*, vol. 54, pp. 2072-2089, May. 2008.

[9] G. V. Moustakides, “Optimal stopping times for detecting changes in distributions,” *Ann. Statist.,* vol.14, pp. 1379-1387, 1986.

[10] E. S. Page, “Continuous inspection schemes,” *Biometrika,* vol. 41, pp. 100-115, 1954.

[11] M. Pollak, “Optimal detection of a change in distribution,” *Ann. Statist.,* vol. 13, pp. 206-227, 1985.

[12] A. N. Shiryayev, “On optimum methods in quickest detection problems,” *Theory Probab. Appl.,* vol. 8, pp. 22-46, 1963.

[13] —, *Optimal Stopping Rules.* New York: Springer-Verlag, 1978.

[14] A. G. Tartakovsky, V. V. Veeravalli, “An efficient sequential procedure for detecting changes in multichannel and distributed systems,” *Proc. 5th Int. Conf. Information Fusion,* vol. 2, Annapolis, MD, pp. 1-8, Jul. 2002.

[15] J. N. Tsitsiklis, “Extremal properties of likelihood ratio quantizers”, *IEEE Trans. Commun.*, vol. 41, pp. 550-558, 1993.

[16] V. V. Veeravalli, “Sequential decision fusion: theory and applications”, *J. Franklin Inst.*, vol. 336, pp. 301-322, Feb. 1999.

[17] —, “Decentralized quickest change detection,” *IEEE Trans. Inf. Theory,* vol. 47, pp. 1657-1665, May 2001.

[18] V. V. Veeravalli, T. Basar, and H. V. Poor, “Decentralized sequential detection with a fusion center performing the sequential test,” *IEEE Trans. Inf. Theory*, vol. 39, pp. 433-442, Mar. 1993.

4.3 Manuscript C: The 2-CUSUM Test

The 2-CUSUM tests (for the Brownian Motion model)

George Moustakides (moustaki@upatras.gr) and Yajun Mei (ymei@isye.gatech.edu)

September 20, 2010

1 Introduction

In this paper we are interested in developing formulas for several performance indexes concerning the two sided CUSUM (2-CUSUM) test applied to the case of the Brownian Motion (BM) process with constant drift. BM models in connection with 2-CUSUM test will become our main model for the problem of of behavioral data modeling treated in detail later. What is really interesting in our approach is that we will come up with closed form expressions for all our performance indexes which then will be used to describe real data in behavior research.

2 The CUSUM test performance

Let $\{\xi\}_{t\geq 0}$ be a BM with constant drift satisfying the following stochastic differential equation (sde)

$$d\xi_t = \mu dt + dw_t \tag{1}$$

where $\{w_t\}_{t\geq 0}$ is a standard Wiener process and μ is the constant drift of the BM. Let us recall the definition and some basic results concerning the CUSUM test

$$\begin{aligned} u_t &= -\frac{\lambda^2}{2}t + \lambda\xi_t + x; \quad x \geq 0 \\ m_t &= \min\{0, \inf_{0\leq s\leq t} u_s\} \\ y_t &= u_t - m_t \\ \mathcal{S} &= \inf\{t \geq 0 : y_t \geq \nu\}. \end{aligned} \tag{2}$$

Process $\{y_t\}_{t\geq 0}$ is known as the CUSUM process which, we observe, *it is started from a nonnegative point x*; $\mathcal{S}$ is the CUSUM stopping time and ν the corresponding threshold. Please note that the CUSUM test defined in (2) with $x = 0$ is optimum when detecting a change in the drift of the BM from 0 to λ [see Shiryaev (1961), Beibel (1961) and Moustakides (2004)]. Here, however, we will allow the CUSUM parameter λ to differ from the actual mean μ of the BM and, furthermore, we are going to allow the test to start from any nonnegative point x instead of the classical case $x = 0$. This will give to our model more flexibility and, also, allow us to infer about the optimality of the human-decision mechanism, if we can verify that the most appropriate model is $\lambda = \mu$.

Using the formulas developed in the literature, we have

$$\phi(x,\mu) = \mathbb{E}^{\mu}[\mathcal{S}|y_0 = x] = \frac{2}{\lambda^2}\frac{e^{\rho\nu} - e^{\rho x} - \rho(\nu - x)}{\rho^2} \tag{3}$$

$$\psi(x,\mu) = \mathbb{E}^{\mu}\left[e^{-s\mathcal{S}}|y_0 = x\right] = \frac{\kappa_2 e^{\kappa_1 x} - \kappa_1 e^{\kappa_2 x}}{\kappa_2 e^{\kappa_1 \nu} - \kappa_1 e^{\kappa_2 \nu}} \tag{4}$$

where

$$\rho = 1 - 2\frac{\mu}{\lambda}; \quad \kappa_i = \frac{1}{2}\left(\rho \pm \sqrt{\rho^2 + \frac{8s}{\lambda^2}}\right); \tag{5}$$

and $\mathbb{E}^{\mu}[\cdot|y_0 = x]$ denotes expectation with respect to the measure induced by the BM defined in (1) for a CUSUM process that starts from $x \geq 0$.

Let us now recall the basic renewal property which we used in order to obtain the previous formulas

Remark 1: In the CUSUM process $y_t, i = 1, 2$ the running minimum m_t whenever it changes it follows u_t. This suggests that whenever m_t changes, we necessarily have

$$y_t = u_t - m_t = 0,$$

or that m_t cannot change outside the set $\{y_t = 0\}$.
This property is very important for the classical CUSUM and it will turn out to be equally crucial for deriving formulas for 2-CUSUM.

3 The 2-CUSUM test

The 2-CUSUM test consists of two one-sided CUSUM tests running in parallel, where one test detects positive changes and the other negative. In this work we are going to consider only the following symmetric case

$$\begin{array}{ll} u_t^1 = -\dfrac{\lambda^2}{2}t + \lambda\xi_t + x_1; \quad x_1 \geq 0 & \qquad u_t^1 = -\dfrac{\lambda^2}{2}t - \lambda\xi_t + x_2; \quad x_2 \geq 0 \\ m_t^1 = \min\{0, \inf\limits_{0\leq s\leq t} u_s^1\} & \qquad m_t^2 = \min\{0, \inf\limits_{0\leq s\leq t} u_s^2\} \\ y_t^1 = u_t^1 - m_t^1 & \qquad y_t^2 = u_t^2 - m_t^2 \\ \mathcal{S}_1 = \inf\{t \geq 0 : y_t^1 \geq \nu\}. & \qquad \mathcal{S}_2 = \inf\{t \geq 0 : y_t^2 \geq \nu\}. \end{array} \tag{6}$$

where $\mathcal{S}_i, i = 1, 2$ are the two one-sided CUSUM branches of the 2-CUSUM test. As we can see, the first component detects positive changes whereas the second negative. Of course the corresponding 2-CUSUM stopping time is the minimum of the two, that is,

$$\mathcal{S} = \mathcal{S}_1 \wedge \mathcal{S}_2. \tag{7}$$

The same test can also be used to make a selection between the positive and the negative case simply by observing which stopping time stops first. If our decision function is $d \in \{1, 2\}$ where "1" denotes decision in favor of the positive change and "2" in favor of the negative, then we can define d as follows

$$d = \begin{cases} 1 & \text{when } \mathcal{S}_1 < \mathcal{S}_2 \\ 2 & \text{when } \mathcal{S}_2 < \mathcal{S}_1 \end{cases} \tag{8}$$

This definition will be used in order to find several interesting performance indexes associated with the 2-CUSUM-based decision mechanism.

3.1 Important 2-CUSUM renewal property

Let us now introduce a very important property enjoyed by 2-CUSUM which was first observed by Siegmund (1985).

Theorem 1. *Let y_t^1, y_t^2 be the two CUSUM statistics of a symmetric 2-CUSUM procedure with initial points x_1, x_2 respectively. Then whenever one of the two statistics reaches a level which is larger than $x_1 + x_2$,* ***for the first time****, the other initializes to 0.*

Proof: Consider the sum process

$$Y_t = y_t^1 + y_t^2 = -\lambda^2 t - m_t^1 - m_t^2.$$

Note that $Y_0 = x_1 + x_2$ and let $\gamma > Y_0$ be a level which is larger than the initial value of the sum process. If one of the two processes reaches the level γ for the first time, then we necessarily have $Y_t \geq \gamma$. In fact we will show that exact equality holds. Let t_0 be the time that Y_t reaches for the first time the value γ. Note that this can happen only during an increase of the process Y_t. During this increase if none of the two processes m_t^1, m_t^2 changes we can see that Y_t will be decreasing with time, contradiction. Hence at least one of the two processes m_t^1, m_t^2 changes at time t_0. Due to Remark 1, if the process m_t^i changes this necessarily implies that $y_t^i = 0$. Note that at t_0 we cannot have both processes changing at the same time since this would imply that $y_{t_0}^1 = y_{t_0}^2 = 0$ or $Y_{t_0} = 0$ which is again contradiction since $Y_{t_0} = \gamma > 0$. Therefore we conclude that *exactly* one process m_t^i must change at t_0. Suppose, for example, that this process is m_t^1 then at time t_0, $y_{t_0}^1 = 0$ and $y_{t_0}^2 = \gamma$, which also suggests that t_0 is the time where one of the process hits the level γ for the first time and the other necessarily initializes to 0.

Based on Theorem 1 we have the following straightforward remark:

Remark 2: Let the symmetric 2-CUSUM procedure be defined as in (6) with initializing points x_1, x_2. If the threshold ν satisfies $\nu > x_1 + x_2$, then when one of the branches stops first, the other branch has CUSUM statistics which is re-initialized to 0.

Remark 2 is crucial in obtaining formulas for the 2-CUSUM test by making use of the corresponding formulas for the one-sided branches.

3.2 2-CUSUM performance

In this subsection, our goal is to find the performance of the 2-CUSUM stopping time with respect to different probability distributions. As before we will assume that the BM process $\{\xi_t\}$ has rate μ and that the two branches start from different nonnegative values x_1, x_2. The formulas we are going to develop, as a result of Theorem 1, will be valid for the case $x_1 + x_2 < \nu$. Unfortunately no formula is known when $x_1 + x_2 \geq \nu$. This poses no particular problem since we are going to consider, mostly, the case $x_1 = x_2 = 0$.

Again we borrow ideas from Siegmund (1985) developed for the discrete-time case and transfer them to the continuous-time continuous-path BM case. The following theorem presents a collection of interesting formulas regarding the symmetric 2-CUSUM.

Theorem 2. *Let $\mathcal{S}$ be the symmetric 2-CUSUM with branches $\mathcal{S}_1, \mathcal{S}_2$, parameters $\pm\lambda$, initial points*

x_1, x_2 and common threshold ν, as in (7). Let $\mathbb{P}^\mu$ be the probability measure and $\mathbb{E}^\mu[\cdot]$ the corresponding expectation induced by the BM process with rate μ. Define the functions $\phi_i(x_i, \mu) = \mathbb{E}^\mu[\mathcal{S}_i | y_0^i = x_i]$ and $\psi_i(x_i, \mu) = \mathbb{E}^\mu[e^{-s\mathcal{S}_i} | y_0^i = x_i]$ referring to the one-sided branches according to (3) and (4). If $x_1 + x_2 < \nu$ then we have the following formulas for 2-CUSUM

$$\mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2] = \frac{\phi_1(x_1,\mu)\phi_2(0,\mu) + \phi_1(0,\mu)\phi_2(x_2,\mu) - \phi_1(0,\mu)\phi_2(0,\mu)}{\phi_1(0,\mu) + \phi_2(0,\mu)} \tag{9}$$

$$\mathbb{P}^\mu[\mathcal{S}_1 < \mathcal{S}_2 | y_0^1 = x_1, y_0^2 = x_2] = \frac{\phi_2(x_2,\mu) + \phi_1(0,\mu) - \phi_1(x_1,\mu)}{\phi_1(0,\mu) + \phi_2(0,\mu)} \tag{10}$$

$$\mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2] = \frac{\psi_1(x_1,\mu)[\psi_2(0,\mu) - 1] + \psi_2(x_2,\mu)[\psi_1(0,\mu) - 1]}{\psi_1(0,\mu)\psi_2(0,\mu) - 1} \tag{11}$$

$$\mathbb{E}^\mu[e^{-s\mathcal{S}_1}\mathbb{1}_{\{\mathcal{S}_1 < \mathcal{S}_2\}} | y_0^1 = x_1, y_0^2 = x_2] = \frac{\psi_1(0,\mu)\psi_2(x_2,\mu) - \psi_1(x_1,\mu)}{\psi_1(0,\mu)\psi_2(0,\mu) - 1}. \tag{12}$$

Proof. The proof borrows ideas from Siegmund (1985) developed for the discrete-time case. We easily verify that

$$\mathcal{S}_1 = \mathcal{S} + [\mathcal{S}_1 - \mathcal{S}_2]\mathbb{1}_{\{\mathcal{S}_1 > \mathcal{S}_2\}}.$$

Applying expectation on both sides of the previous equality we can write

$$\begin{aligned}\phi_1(x_1,\mu) &= \mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2] + \mathbb{E}^\mu\left[\mathbb{E}^\mu[\mathcal{S}_1 - \mathcal{S}_2 | \mathcal{F}_{\mathcal{S}_2}]\mathbb{1}_{\{\mathcal{S}_1 > \mathcal{S}_2\}} | y_0^1 = x_1, y_0^2 = x_2\right] \\ &= \mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2] + \phi_1(0,\mu)\mathbb{P}^\mu\left[\mathcal{S}_1 > \mathcal{S}_2 | y_0^1 = x_1, y_0^2 = x_2\right]. \end{aligned} \tag{13}$$

Note that going from the first equation to the second we used the fact that since we consider the event $\mathcal{S}_2 < \mathcal{S}_1$, we have that at time $\mathcal{S}_2$ the first CUSUM branch restarts, therefore $\mathbb{E}^\mu[\mathcal{S}_1 - \mathcal{S}_2 | \mathcal{F}_{\mathcal{S}_2}] = \phi_1(0,\mu)$. The same way we can show that

$$\phi_2(x_2,\mu) = \mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2] + \phi_2(0,\mu)\mathbb{P}^\mu\left[\mathcal{S}_2 > \mathcal{S}_1 | y_0^1 = x_1, y_0^2 = x_2\right]. \tag{14}$$

Solving (13) for $\mathbb{P}^\mu\left[\mathcal{S}_1 > \mathcal{S}_2 | y_0^1 = x_1, y_0^2 = x_2\right]$ and (14) for $\mathbb{P}^\mu\left[\mathcal{S}_2 > \mathcal{S}_1 | y_0^1 = x_1, y_0^2 = x_2\right]$, we obtain

$$\mathbb{P}^\mu\left[\mathcal{S}_1 > \mathcal{S}_2 | y_0^1 = x_1, y_0^2 = x_2\right] = \frac{\phi_1(x_1,\mu) - \mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2]}{\phi_1(0,\mu)} \tag{15}$$

$$\mathbb{P}^\mu\left[\mathcal{S}_2 > \mathcal{S}_1 | y_0^1 = x_1, y_0^2 = x_2\right] = \frac{\phi_2(x_2,\mu) - \mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2]}{\phi_2(0,\mu)}. \tag{16}$$

Adding the two equations and solving for $\mathbb{E}^\mu[\mathcal{S} | y_0^1 = x_1, y_0^2 = x_2]$ yields (9). Replacing (9) in (15) we obtain (10).

For proving the other two relations we proceed similarly. In particular we observe that

$$e^{-s\mathcal{S}_1} = e^{-s\mathcal{S}} + \left[e^{-s\mathcal{S}_1} - e^{-s\mathcal{S}_2}\right] \mathbb{1}_{\{\mathcal{S}_1 > \mathcal{S}_2\}} = e^{-s\mathcal{S}} + \left[e^{-s(\mathcal{S}_1 - \mathcal{S}_2)} - 1\right] e^{-s\mathcal{S}_2} \mathbb{1}_{\{\mathcal{S}_1 > \mathcal{S}_2\}}.$$

Taking expectation of the previous equality we obtain

$$\psi_1(x_1, \mu) = \mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2] + [\psi_1(0, \mu) - 1]\mathbb{E}^\mu\left[e^{-s\mathcal{S}_2}\mathbb{1}_{\{\mathcal{S}_1 > \mathcal{S}_2\}} | y_0^1 = x_1, y_0^2 = x_2\right], \quad (17)$$

and in the same way

$$\psi_2(x_2, \mu) = \mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2] + [\psi_2(0, \mu) - 1]\mathbb{E}^\mu\left[e^{-s\mathcal{S}_1}\mathbb{1}_{\{\mathcal{S}_2 > \mathcal{S}_1\}} | y_0^1 = x_1, y_0^2 = x_2\right]. \quad (18)$$

Solving for the last term in both equation and adding the two equalities we obtain

$$\begin{aligned}\mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2] =& \frac{\psi_1(x_1, \mu) - \mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2]}{\psi_1(0, \mu) - 1} \\ &+ \frac{\psi_2(x_1, \mu) - \mathbb{E}^\mu[e^{-s\mathcal{S}} | y_0^1 = x_1, y_0^2 = x_2]}{\psi_2(0, \mu) - 1}.\end{aligned}$$

Solving for the desired quantity yields (11). Substituting (11) in (18) yields (12).

Theorem 2 will become the basis for finding formulas for even more exotic quantities. Such possibilities could be the pdf of the 2-CUSUM stopping time or the conditional pdf of $\mathcal{S}_1$ given that $\mathcal{S}_1 < \mathcal{S}_2$. Of course this amounts to Laplace-inverting the two functions in (11) and (12), a task which is not very straightforward, as we will find out next.

4 Performance computation

As we can see from Theorem 2, performance of 2-CUSUM depends on the drift value μ of the BM. We are going to consider two special case. In the first we will compute the performance of the test when $\mu = 0$ and in the second $\mu = \lambda$, i.e. the drift equal to the CUSUM parameter parameter. Let us consider each case separately, since the corresponding results tend to be quite different in nature.

4.1 Case of zero drift

The first step is to compute the performance of the two branches. Note that the two branches differ in their CUSUM parameter since the first uses $+\lambda$ and the second $-\lambda$. Considering the case $\mu = 0$, we have from (5) that for both branches

$$\rho = 1, \quad \kappa_j = \frac{1}{2}\left(1 \pm \sqrt{1 + \frac{8s}{\lambda^2}}\right), \; j = 1, 2.$$

This immediately translates into the following expressions for the two branches

$$\begin{aligned}\phi_1(x,0) = \phi_2(x,0) &= \frac{2}{\lambda^2}\left\{e^{\nu} - e^{x} - (\nu - x)\right\}\\ \psi_1(x,0) = \psi_2(x,0) &= \frac{\kappa_2 e^{\kappa_1 x} - \kappa_1 e^{\kappa_2 x}}{\kappa_2 e^{\kappa_1 \nu} - \kappa_1 e^{\kappa_2 \nu}}.\end{aligned}$$

If we now assume that the two branches start from $x = 0$ then

$$\begin{aligned}\phi_1(0,0) = \phi_2(0,0) &= \frac{2}{\lambda^2}\left\{e^{\nu} - 1 - \nu\right\}\\ \psi_1(0,0) = \psi_2(0,0) &= \frac{\kappa_2 - \kappa_1}{\kappa_2 e^{\kappa_1 \nu} - \kappa_1 e^{\kappa_2 \nu}}.\end{aligned}$$

We can now compute the performance of the 2-CUSUM test for $x_1 = x_2 = 0$. Since $\phi_1(x,0) = \phi_2(x,0)$ and $\psi_1(x,0) = \psi_2(x,0)$, we have the simplified expressions

$$\mathbb{E}^0[\mathcal{S}] = \frac{1}{2}\phi_1(0,0) = \frac{1}{\lambda^2}\{e^{\nu} - \nu - 1\} \tag{19}$$

$$\mathbb{P}^0[\mathcal{S}_1 > \mathcal{S}_2] = \frac{1}{2} \tag{20}$$

$$\mathbb{E}^0[e^{-s\mathcal{S}}] = 2\frac{\psi_1(0,0)}{\psi_1(0,0)+1} = 2\frac{\kappa_2 - \kappa_1}{\kappa_2 e^{\kappa_1 \nu} - \kappa_1 e^{\kappa_2 \nu} + \kappa_2 - \kappa_1} = 2\frac{1}{\dfrac{\kappa_2 e^{\kappa_1 \nu} - \kappa_1 e^{\kappa_2 \nu}}{\kappa_2 - \kappa_1} + 1} \tag{21}$$

$$\mathbb{E}^0[e^{-s\mathcal{S}_1}\mathbb{1}_{\{\mathcal{S}_1 < \mathcal{S}_2\}}] = \frac{\psi_1(0,0)}{\psi_1(0,0)+1} \Rightarrow \mathbb{E}^0[e^{-s\mathcal{S}_1} | \mathcal{S}_1 < \mathcal{S}_2] = \mathbb{E}^0[e^{-s\mathcal{S}}]. \tag{22}$$

One of the most challenging problems in this analysis is finding the pdf of the stopping time $\mathcal{S}$, that is $\mathbb{P}^0[\mathcal{S} \in dt]$ or the conditional pdf $\mathbb{P}^0[\mathcal{S}_1 \in dt | \mathcal{S}_1 < \mathcal{S}_2]$. Obtaining the pdf $q^0(t)$ of the 2-CUSUM stopping time, amounts to Laplace-inverting the moment generating function $\mathbb{E}^0[e^{-s\mathcal{S}}]$.

4.1.1 Inverse Fourier transform computation of the pdf

As we said, in order to find the probability density $q^0(t)$ of $\mathcal{S}$, we need to Laplace-invert the function $\mathbb{E}^0[e^{-s\mathcal{S}}]$ with respect to the variable s. Clearly $q^0(t)$ is a *causal* function of time t, that is, the support of $q^0(t)$ is $[0, +\infty)$. Furthermore $q^0(t)$ is absolutely integrable since by being a pdf we have $|q^0(t)| = q^0(t)$ and $\int_0^\infty q^0(t)\,dt = 1$. It is known that causal functions that are absolutely integrable have Laplace transforms that converge for s in the complex right half plane $\mathrm{Re}\{s\} \geq 0$. Consequently, if we call

$$\mathcal{Q}^0(s) = \mathbb{E}^0[e^{-s\mathcal{S}}] = \int_0^\infty e^{-st} q^0(t)\, dt,$$

then $\mathcal{Q}^0(s)$ has all its poles in the left half plane (excluding the origin). This means that we can apply the Laplace inversion formula and integrate along the path $\mathrm{Re}\{s\} = 0$ as follows

$$q^0(t) = \frac{1}{2\pi j} \int_{0-j\infty}^{0+j\infty} \mathcal{Q}^0(s) e^{st}\, ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{Q}^0(j\Omega) e^{j\Omega t}\, d\Omega, \tag{23}$$

with the last integral interpreted as an inverse Fourier transform and obtained by replacing s with $j\Omega$. The inverse Fourier transform formula can lend itself to a numerical computation of the pdf

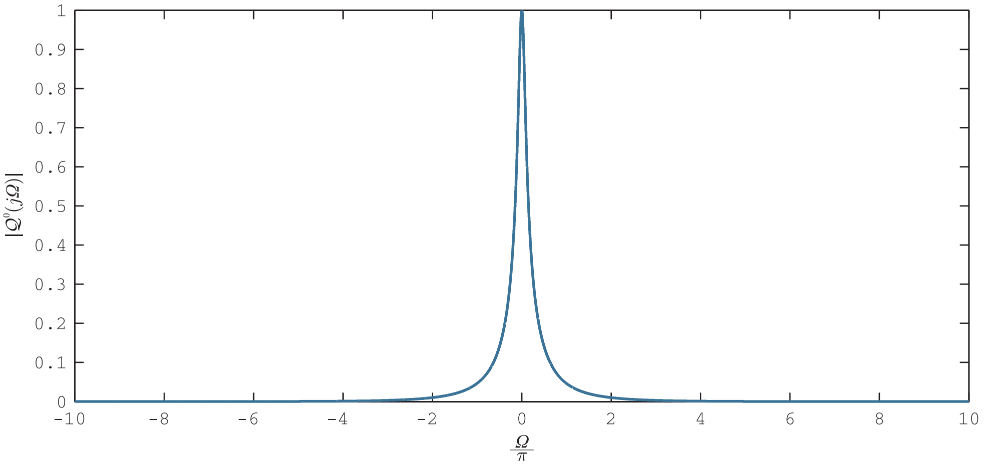

Figure 1: Typical form of the characteristic function $|\mathcal{Q}^0(j\Omega)|$ of the 2-CUSUM stopping time.

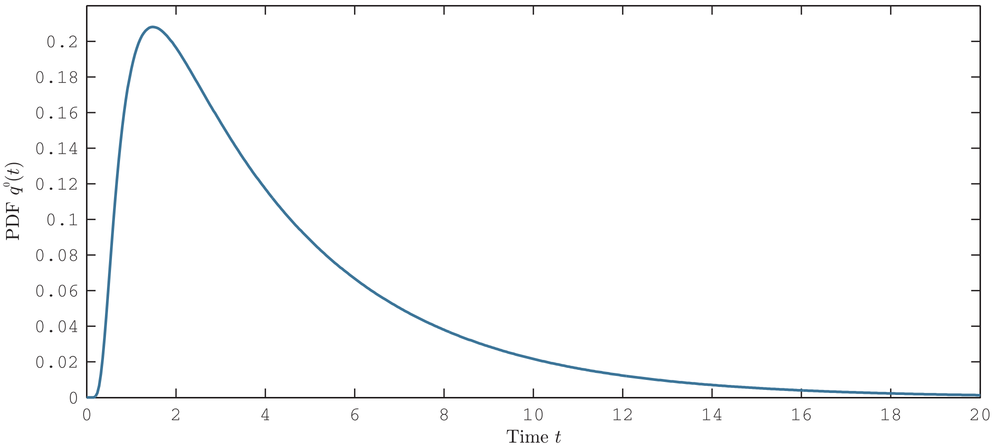

Figure 2: Typical form of the 2-CUSUM stopping time pdf function $q^0(0)$.

function $q^0(t)$. We can see in Fig. 1 the typical form of the amplitude of the function $\mathcal{Q}^0(j\Omega)$ for the case $\lambda = 1$ and $\nu = 2$. The pdf $q^0(t)$, it is known that, when $\mu = 0$, has exponential tails. Indeed we can see in Fig. 2 this fact for a threshold $\nu = 2$. Tail behavior is usually an important issue for random variables. Therefore in the next part we will follow an alternative path to describe the pdf $q^0(t)$, in which tail behavior will be more apparent.

4.1.2 Series expansion of the pdf

Another approach for computing the pdf $q^0(t)$ consists in expanding the moment generating function $\mathcal{Q}^0(s)$ as a series

$$\mathcal{Q}^0(s) = \sum_{k=1}^{\infty} \frac{\mathcal{A}_k}{s - s_k}; \quad \text{where } \mathcal{A}_k = \lim_{s \to s_k} (s - s_k)\mathcal{Q}^0(s).$$

Sequence $\{s_k\}$ is the collection of poles of the function $\mathcal{Q}^0(s)$ and $\{\mathcal{A}_k\}$, as we can see, the corresponding sequence of residues of the poles. Please note that, for simplicity, we have assumed that the poles have single multiplicity which is true when the previous limit is finite. Such an expansion, when Laplace inverted, yields

$$q^0(t) = \sum_{k=1}^{\infty} \mathcal{A}_k e^{s_k t}, \tag{24}$$

Equ. (24) can help in identifying the exponential tail of $q^0(t)$ since this will simply correspond[1] to the term $A_1 e^{s_1 t}$, assuming that we have ranked the poles in decreasing order regarding their real parts, that is, $\mathrm{Re}\{s_1\} > \mathrm{Re}\{s_2\} > \cdots$, and recalling that these real parts are negative.

According to (24) we first need to find the poles of the function $\mathcal{Q}^0(s)$ defined in (21). Unfortunately, as we will see shortly, this can be achieved only numerically. However, asymptotically as $\nu \to \infty$, we will be able to describe this exponential decay in a more accurate analytical way.

Define the following function of x

$$\mathcal{V}(x) = \frac{(1+x)e^{0.5(1-x)\nu} - (1-x)e^{0.5(1+x)\nu}}{2x} = e^{0.5\nu}\left\{\cosh(0.5\nu x) - x^{-1}\sinh(0.5\nu x)\right\} \tag{25}$$

then

$$\mathcal{Q}^0(s) = \frac{2}{\mathcal{V}\left(\sqrt{1 + \frac{8s}{\lambda^2}}\right) + 1}.$$

From the previous relation we conclude that if $\{x_k\}$ is the collection of roots of the equation $\mathcal{V}(x) + 1 = 0$ then the poles of $\mathcal{Q}^0(s)$ are simply $s_k = (x_k^2 - 1)\lambda^2/8,\ k = 1, 2, \ldots$. We can also find

[1] Of course we need to prove absolute summability of the remaining terms.

for these poles the corresponding residues in terms of x_k. Note that if we call $x = \sqrt{1 + 8s/\lambda^2}$ then $s = (x^2 - 1)\lambda^2/8$ and clearly $s \to s_k$ means that $x \to x_k$. Therefore we can write

$$\mathcal{A}_k = \lim_{s \to s_k} (s - s_k)\mathcal{Q}^0(s) = \frac{\lambda^2}{4} \lim_{x \to x_k} \frac{x^2 - x_k^2}{\mathcal{V}(x) + 1} = \frac{\lambda^2}{4} 2x_k \lim_{x \to x_k} \frac{x - x_k}{\mathcal{V}(x) + 1} = \frac{\lambda^2}{2} \frac{x_k}{\mathcal{V}'(x_k)}, \tag{26}$$

where $\mathcal{V}'(x)$ denotes derivative. In the last equation we made use of Hospital's rule since $\mathcal{V}(x_k)+1 = 0$ and we assumed that $\mathcal{V}'(x_k) \neq 0$. From this and (24) we obtain the following series expansion for $q^0(t)$

$$q^0(t) = \frac{\lambda^2}{2} \sum_{k=1}^{\infty} \frac{x_k}{\mathcal{V}'(x_k)} e^{\frac{\lambda^2}{8}(x_k^2 - 1)t}, \ t \geq 0. \tag{27}$$

To complete the determination of the pdf $q^0(t)$, we need to find the roots $\{x_k\}$ of the equation $\mathcal{V}(x) + 1 = 0$. We have the following *conjecture* for this problem.

Conjecture 1. *Regarding the poles of $\mathcal{Q}^0(s)$ defined in (21), we have the following claims:*

i) All poles of $\mathcal{Q}^0(s)$ are **real** *and negative.*

ii) When $\nu \geq 2.5569$ then $\mathcal{Q}^0(s)$ has a pole in the interval $[-\lambda^2/8, 0)$ and all other poles (infinite number) in the interval $(-\infty, -\lambda^2/8)$.

iii) When $0 \leq \nu < 2.5569$, then $\mathcal{Q}^0(s)$ has all its poles in the interval $(-\infty, -\lambda^2/8]$.

We note that the function $\mathcal{V}(x)$ is symmetric in x. This means that if x_k is a root of $\mathcal{V}(x) + 1 = 0$, so is $-x_k$. Since both values produce the same pole s_k, we are going to limit ourselves to $x \geq 0$. Let us first seek the real solutions of the equation $\mathcal{V}(x) + 1$. Since for $z \geq 0$ we have $\cosh(z) > \sinh(z)$ we conclude that $\mathcal{V}(x) + 1 > 0$ when $x > 1$. Therefore if a real root exists, it has to lie inside the interval [0,1]. We note that $\mathcal{V}(1) + 1 = 2 > 0$ while $\mathcal{V}(0) + 1 = (1 - 0.5\nu)e^{0.5\nu} + 1$. The latter value is nonpositive when $\nu \geq 2.5569$ assuring existence of a root inside [0,1]. We need to show that this root is unique when ν is less than 2.5569 and nonexistent when ν is smaller than this value. In Fig.3 we see the two forms of the function $\mathcal{V}(x) + 1$ for the two possibilities of ν. This root x_1 corresponds to a real pole s_1 in the interval $[-\lambda^2/8, 0)$.

Let us now consider imaginary solutions of $\mathcal{V}(x) + 1 = 0$. By replacing $x = j\omega$, the equation is equivalent to

$$A(\omega) = e^{-0.5\nu}\{\mathcal{V}(j\omega) + 1\} = \cos(0.5\nu\omega) - \omega^{-1}\sin(0.5\nu\omega) + e^{-0.5\nu} = 0. \tag{28}$$

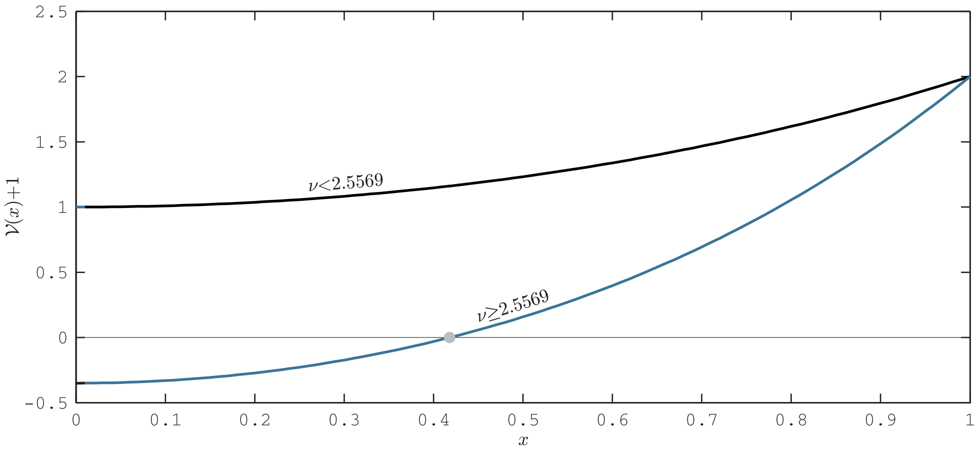

Figure 3: Form of the function $\mathcal{V}(x)+1$ for $\nu \geq 2.5569$ (blue) and $\nu < 2.5569$ (black).

where the hyperbolic functions are transformed into regular trigonometric functions. Fig. 4 depicts a typical form of the function $A(\omega)$. We can see that there is an infinite number of roots. In fact as ω grows, the roots approach the roots of the equation $\cos(0.5\nu\omega) + e^{-0.5\nu} = 0$ which are easy to compute.

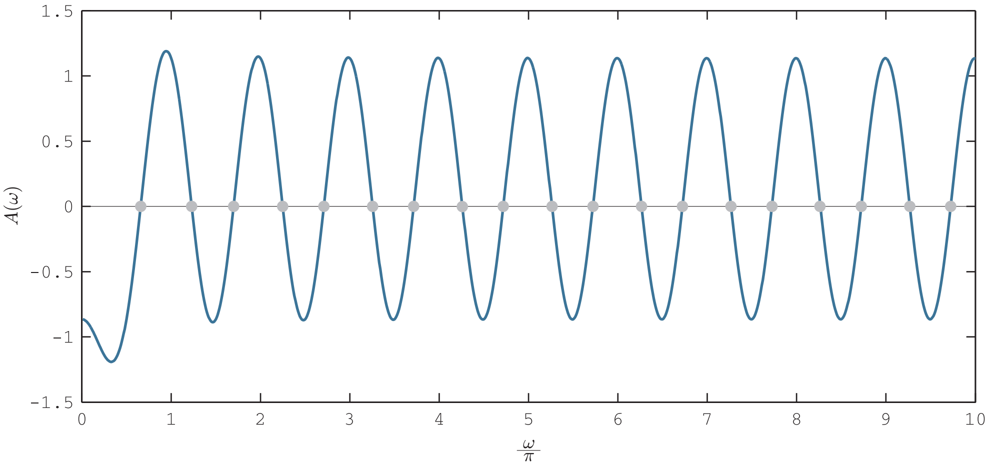

Figure 4: Typical form of the function $A(\omega)$ and corresponding sequence of roots $\{\omega_k\}$ (gray circles).

Of course the actual computation of the roots of the equation $A(\omega) = 0$ is important for applying the series expansion formula in (27). However, if we are interested in describing the tails of $q^0(t)$ we need to specify the *smallest* pole s_1. We note that whenever $\mathcal{V}(x)+1$ has a real root, which surely happens when $\nu \geq 2.5569$ then this generates the smallest pole. Indeed if we call this root

x_1 then the corresponding pole is $s_1 = (x_1^2 - 1)\lambda^2/8$ and since $x_1 \in [0, 1)$ we have $-\lambda^2/8 \leq s_1 < 0$. The other imaginary roots of the form $x_k = j\omega_k$ give rise to poles of the form

$$s_k = -(\omega_k^2 + 1)\frac{\lambda^2}{8} < -\frac{\lambda^2}{8} \leq s_1 < 0,$$

suggesting that s_1 is the leading pole that defines the exponential tail behavior of the pdf $q^0(t)$. Of course when $0 < \nu < 2.5569$ then, as we said, there is no real root for $\mathcal{V}(x) + 1$ and therefore the tail behavior is governed by the smallest, in amplitude, imaginary root $x_k = j\omega_k$.

As an example let us compute the roots for the case $\nu = 2$ and the corresponding pdf $q^0(t)$ using the series expansion (27). From the above we know that we have only imaginary roots x_k. The smallest in amplitude is $x_1 = j1.1198$. The next roots have the values $j4.1081$, $j8.1050$, $j10.5259$, $j14.4438$, $j16.8434$, $\ldots$, while the poles s_k become -0.2818, -2.2345, -8.3363, -13.9743, -26.2030, -35.5876, $\ldots$ with corresponding residues $\mathcal{A}_k$ equal to 0.3604, -0.8281, 1.6147, -2.0894, 2.8633, -3.3365, $\ldots$. If we compare the pdf $q^0(t)$ obtained using the series expansion (27) with the numerical integration using (23) the two curves are indistinguishable, therefore the graph is the same as in (2). What is more interesting however is to plot the series expansion using a logarithmic scale for $q^0(t)$ and compare it to the leading exponential term $\mathcal{A}_1 e^{s_1 t}$. In Fig, 5 we can see the two curves. We observe that after some point (in fact very rapidly), the leading exponential term (black) prevails, taking over the tails (blue) completely. This phenomenon becomes more pronounced if we consider large values for the threshold ν. It is exactly this case we would like to consider in the sequel.

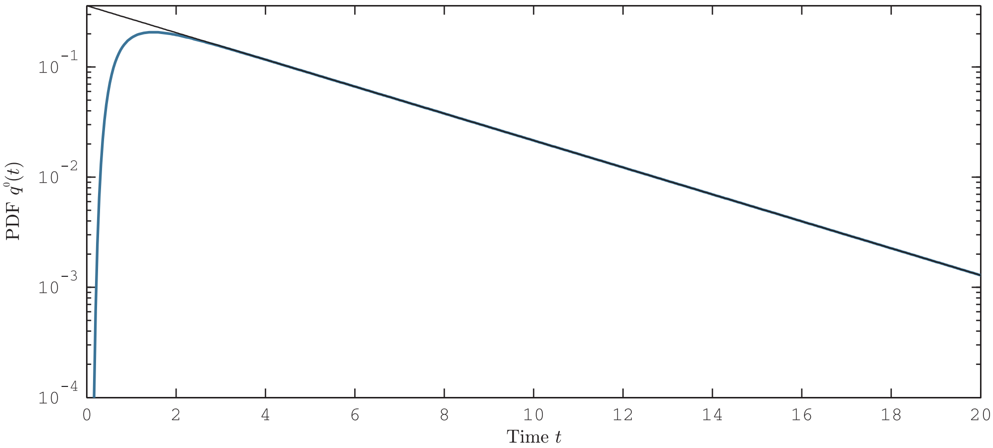

Figure 5: Comparison of pdf $q^0(t)$ (blue) and leading exponential term $\mathcal{A}_1 e^{s_1 t}$ (black).

4.1.3 Asymptotic performance for large thresholds

In this part we would like to analyze the asymptotic performance of the leading exponential term for threshold values that satisfy $\nu \gg 1$. We must point out that our reasoning *is not going to be mathematically rigorous.* We will only sketch the main idea of the proof. Let us begin by considering the real root x_1 of the equation $\mathcal{V}(x)+1$. For large values of ν we observe that $\mathcal{V}(1)+1=2$ while $\mathcal{V}(0)+1=O(e^{0.5\nu}) \gg 2$ which suggests that x_1 must be closer to 1 than to 0. Assuming that x_1 is a (negative) perturbation of 1, that is, $x_1 = 1-\epsilon$ where $0<\epsilon \ll 1$ and linearizing the function $\mathcal{V}(x)+1$ around $x=1$, we find

$$\epsilon = 4e^{-\nu}(1+o(1)).$$

This implies that the first pole is extremely small and equal to $s_1 = -\lambda^2 e^{-\nu}(1+o(1))$. If we substitute this value in the formula (26), we obtain $\mathcal{A}_1 = \lambda^2 e^{-\nu}(1+o(1))$ which is of the same order as s_1.

For the remaining imaginary solutions of the equation, we can see that for large ν and large ω, the function $A(\omega)$ in (28) behaves like $\cos(0.5\nu\omega)$, which suggests that $\omega_k \approx (2k-3)O(1)/\nu,\ k=2,3,\dots$ yielding poles $s_k = -\lambda^2[1+(2k-3)^2O(1)/\nu^2]/8$ and residues $\mathcal{A}_k = \lambda^2 e^{-0.5\nu}(-1)^k(2k-3)O(1)/\nu^2,\ k=2,3,\dots$. In both cases the term $O(1)$ is uniform over k and ν.

We can now verify the correctness of these values by computing the average detection delay. We have that

$$\begin{aligned}\int_0^\infty tq^0(t)\,dt &= \sum_{k=1}^\infty \frac{A_k}{s_k^2} = \frac{A_1}{s_1^2} + \sum_{k=2}^\infty \frac{A_k}{s_k^2} \\ &= \frac{1}{\lambda^2}\left\{e^{\nu}(1+o(1)) + e^{-0.5\nu}\nu^2 \sum_{k=2}^\infty \frac{(-1)^k(2k-3)O(1)}{[\nu^2+(2k-3)^2O(1)]^2}\right\}.\end{aligned}$$

This suggests that

$$\begin{aligned}\left|\int_0^\infty tq^0(t)\,dt - \frac{A_1}{s_1^2}\right| &\le \frac{1}{\lambda^2}\left\{e^{-0.5\nu}\nu^2 O(1) \sum_{k=0}^\infty \frac{1}{(2k+1)^3}\right\} \\ &= \frac{1}{\lambda^2} e^{-0.5\nu}\nu^2 O(1) \to 0.\end{aligned}$$

For the last equality we used the fact that $\sum_{k=0}^\infty 1/(2k+1)^3 < \infty$ and the limit is for $\nu \to \infty$. We can therefore see that the difference between the average time and the average of the leading exponential, tends to zero as the threshold increases, even though both averages tend to infinity.

We therefore conclude that $\mathbb{E}^0[\mathcal{S}] = e^{\nu}(1+o(1))/\lambda^2$ and this conclusion is corroborated by the exact formula for $\mathbb{E}^0[\mathcal{S}]$ in (19). Since we have specified a rate of convergence of $\mathcal{A}_1/s_1^2$ towards the mean time, i.e. $O(e^{-0.5\nu}\nu^2)$, it is clear that we can use higher order estimates for $\mathcal{A}_1, s_1$, up to and including $O(1)$ terms. These terms can be estimated accurately since they are larger, in order of magnitude, than the convergence rate.

Regarding the form of the pdf under large values of the threshold ν, by comparing the poles, we arrive at the following important conclusion:

Remark 3: As the threshold $\nu \to \infty$, we note a clear separation between the leading pole that tends to 0 and all other poles that tend to $-\lambda^2/8$. This guarantees an exponential tail behavior even in the limit, as $\nu \to \infty$.

It is interesting to prove this remark formally. Borrowing ideas from Taylor (1975), we will show that, as $\nu \to \infty$, a properly normalized version of the stopping time $\mathcal{S}$ tends to a pure exponential. Indeed consider the following version of the stopping time $\tilde{\mathcal{S}} = e^{-\nu}\mathcal{S}$. Then

$$\mathbb{E}^0[e^{-s\tilde{\mathcal{S}}}] = \mathbb{E}^0[e^{-(se^{-\nu})\mathcal{S}}] = \mathcal{Q}^0(se^{-\nu}).$$

Fix s, with $\mathrm{Re}\{s\} \geq 0$, and consider the limit

$$\lim_{\nu\to\infty} \mathbb{E}^0[e^{-s\tilde{\mathcal{S}}}] = \lim_{\nu\to\infty} \mathcal{Q}^0(se^{-\nu})$$

Note that we can write $\sqrt{1+\epsilon} = 1 + 0.5\epsilon + o(\epsilon)$ and $(1+\epsilon)^{-1/2} = 1 - 0.5\epsilon + o(\epsilon)$ which, if used in the previous limit after observing that $|8se^{-\nu}/\lambda^2| \ll 1$, we can write

$$\begin{aligned}\lim_{\nu\to\infty} \mathbb{E}^0[e^{-s\tilde{\mathcal{S}}}] &= \lim_{\nu\to\infty} \frac{2}{\mathcal{V}\left(\sqrt{1+\frac{8se^{-\nu}}{\lambda^2}}\right)+1} \\ &= \lim_{\nu\to\infty} \frac{2}{e^{0.5\nu}[e^{-0.5\nu\{1+O(e^{-\nu})\}} + e^{0.5\nu\{1+O(e^{-\nu})\}}(2se^{-\nu}/\lambda^2)]+1} \\ &= \frac{1}{1+\frac{s}{\lambda^2}}.\end{aligned}$$

The latter result suggests that, in the limit, $e^{-\nu}\mathcal{S}$ is exponentially distributed with mean $1/\lambda^2$, which of course corresponds to the leading pole of $\mathcal{Q}^0(s)$. Consequently, for large values of the threshold, the leading exponential term prevails completely over the other terms. As we are going to see next, this is not at all the case when the drift of the process $\{\xi_t\}$ is not equal to 0.

4.2 Case of drift equal to the CUSUM parameter

Let us now consider the case where the BM drift μ is equal to the CUSUM parameter λ. Here the first branch with parameter λ must detect the change whereas the other branch with parameter $-\lambda$ should not. In other words $\mathcal{S}_1 < \mathcal{S}_2$ corresponds to a correct decision while the opposite to a wrong one. Of course in this case the error probability is expected to be (much) less than 0.5.

Let us again compute the necessary one-sided CUSUM expressions. We start by computing the functions $\phi_1(x,\lambda)$, $\phi_2(x,\lambda)$, $\psi_1(x,\lambda)$, $\psi_2(x,\lambda)$. Note that for $\mu = \lambda$ we have $\rho = -1$ and $\kappa_{1,2} = 0.5(-1 \pm \sqrt{1+8s/\lambda^2})$ which yields

$$\phi_1(x,\lambda) = \frac{2}{\lambda^2}\left\{\nu - x + e^{-\nu} - e^{-x}\right\} \tag{29}$$

$$\psi_1(x,\lambda) = \frac{(-1+\sqrt{1+8s/\lambda^2})e^{-0.5(1+\sqrt{1+8s/\lambda^2})x} + (1+\sqrt{1+8s/\lambda^2})e^{0.5(-1+\sqrt{1+8s/\lambda^2})x}}{(-1+\sqrt{1+8s/\lambda^2})e^{-0.5(1+\sqrt{1+8s/\lambda^2})\nu} + (1+\sqrt{1+8s/\lambda^2})e^{0.5(-1+\sqrt{1+8s/\lambda^2})\nu}}. \tag{30}$$

Similarly for the second one-sided CUSUM branch, by considering $\mu = \lambda$ and CUSUM parameter $-\lambda$ we obtain $\rho = 3$, $\kappa_{1,2} = 0.5(3 \pm \sqrt{9+8s/\lambda^2})$. Thus

$$\phi_2(x,\lambda) = \frac{2}{9\lambda^2}\{e^{3\nu} - e^{3x} - 3(\nu - x)\} \tag{31}$$

$$\psi_2(x,\lambda) = \frac{(3+\sqrt{9+8s/\lambda^2})e^{0.5(3-\sqrt{9+8s/\lambda^2})x} - (3-\sqrt{9+8s/\lambda^2})e^{0.5(3+\sqrt{9+8s/\lambda^2})x}}{(3+\sqrt{9+8s/\lambda^2})e^{0.5(3-\sqrt{9+8s/\lambda^2})\nu} - (3-\sqrt{9+8s/\lambda^2})e^{0.5(3+\sqrt{9+8s/\lambda^2})\nu}}. \tag{32}$$

Applying the general relations we have developed in Theorem 2 and assuming that both initializing points satisfy $x_1 = x_2 = 0$, we end up with the following equations

$$\mathbb{E}^\lambda[\mathcal{S}] = \frac{\phi_1(0,\lambda)\phi_2(0,\lambda)}{\phi_1(0,\lambda) + \phi_2(0,\lambda)} \tag{33}$$

$$\mathbb{P}^\lambda[\mathcal{S}_1 > \mathcal{S}_2] = \frac{\phi_1(0,\lambda)}{\phi_1(0,\lambda) + \phi_2(0,\lambda)} \tag{34}$$

$$\mathbb{E}^\lambda[e^{-s\mathcal{S}}] = \frac{2\psi_1(0,\lambda)\psi_2(0,\lambda) - \psi_1(0,\lambda) - \psi_2(0,\lambda)}{\psi_1(0,\lambda)\psi_2(0,\lambda) - 1} \tag{35}$$

$$\mathbb{E}^\lambda[e^{-s\mathcal{S}_1}\mathbb{1}_{\{\mathcal{S}_1 < \mathcal{S}_2\}}] = \frac{\psi_1(0,\lambda)[\psi_2(0,\lambda) - 1]}{\psi_1(0,\lambda)\psi_2(0,\lambda) - 1}. \tag{36}$$

We can see that as far as the $\phi_i(0,\lambda)$ functions are concerned, for threshold $\nu \gg 1$, we have that

$$\phi_2(0,\lambda) \gg \phi_1(0,\lambda).$$

Indeed we observe that $\phi_1(0,\lambda)$ increases linearly with ν whereas $\phi_2(0,\lambda)$ exponentially. This has the following consequence

$$\mathbb{E}^{\lambda}[\mathcal{S}] = \frac{2}{\lambda^2}\{\nu - 1 + o(1)\} \tag{37}$$

$$\mathbb{P}^{\lambda}[\mathcal{S}_1 > \mathcal{S}_2] = \frac{1}{9}e^{-3\nu}\{\nu - 1 + o(1)\} = \frac{1}{9}e^{-3\nu}\nu(1 + o(1)). \tag{38}$$

In Figs. 6 and 7 we plot the exact value of the two quantities as a function of ν for the case $\lambda = 1$. We observe a rapid convergence to the linear and exponential part respectively.

4.2.1 Empirical formula

In the literature the empirical speed-accuracy tradeoff formula has been observed and reported, see Wickelgren(1977) and Zhang and Chang (2008):

$$P(\text{correct decision}) = 1 - \exp(-a(t - b)), \qquad (t \geq b > 0, a > 0). \tag{39}$$

Equivalently, under our notations, the probability $\mathbb{P}^{\lambda}[\mathcal{S}_1 > \mathcal{S}_2]$ of deciding erroneously is proportional to $\mathbb{E}^{\lambda}[e^{-s\mathcal{S}}]$ for properly selected exponent s. In other words it has been experimentally verified that the following relation is true

$$\text{Prob}\,\{\text{Wrong decision}\} = \mathbb{P}^{\lambda}[\mathcal{S}_1 > \mathcal{S}_2] = \mathcal{C} \times \mathbb{E}^{\lambda}[e^{-s\mathcal{S}}],$$

where $s = a$ and $\mathcal{C} = \exp(ab)$ in view of relation (39).

As we are going to see, our analysis supports a slightly different expression

$$\text{Prob}\,\{\text{Wrong decision}\} = \mathbb{P}^{\lambda}[\mathcal{S}_1 > \mathcal{S}_2] = \mathcal{C} \times \mathbb{E}^{\lambda}[\mathcal{S}] \times \mathbb{E}^{\lambda}[e^{-s\mathcal{S}}]$$

We must emphasize that the extra term $\mathbb{E}^{\lambda}[\mathcal{S}] = O(\nu)$ is not really important since, from an asymptotic point of view, its contribution to the exponential term $\mathbb{E}^{\lambda}[e^{-s\mathcal{S}}]$ is negligible.

Adopting this new model, let us attempt to specify analytically s and $\mathcal{C}$ considering that we are in the asymptotic case $\nu \gg 1$. Note that $\mathbb{E}^{\lambda}[e^{-s\mathcal{S}}]$ is given in (35). Assuming that s is real and positive, this immediately suggests the following approximations

$$\begin{aligned}
\mathbb{E}^{\lambda}[\mathcal{S}] &= \frac{2}{\lambda^2}\nu(1 + o(1))\\
\psi_1(0,\lambda) &= \frac{2\sqrt{1 + 8s/\lambda^2}}{1 + \sqrt{1 + 8s/\lambda^2}}e^{-0.5(\sqrt{1+8s/\lambda^2}-1)\nu}(1 + o(1))\\
\psi_2(0,\lambda) &= \frac{2\sqrt{9 + 8s/\lambda^2}}{\sqrt{9 + 8s/\lambda^2} - 1}e^{-0.5(3+\sqrt{9+8s/\lambda^2})}(1 + o(1)).
\end{aligned}$$

From the previous formulas we realize that $1 \gg \psi_1(0,\lambda) \gg \psi_2(0,\lambda)$ which if used in (35) yields

$$\mathbb{E}^{\lambda}[e^{-s\mathcal{S}}] = \{\psi_1(0,\lambda) + \psi_2(0,\lambda)\}(1 + o(1)) = \psi_1(0,\lambda)(1 + o(1)). \tag{40}$$

In order to specify the two parameters $\mathcal{C}$ and s we need to equate (40) to (38). Equating first the exponential parts and in particular the two exponents, results in the following relation for s

$$0.5(\sqrt{1 + 8s/\lambda^2} - 1) = 3$$

which yields

$$s = 6 \times \lambda^2.$$

The proportionality constant $\mathcal{C}$ we are looking for, can now be computed as the limit

$$\mathcal{C} = \lim_{\nu \to \infty} \frac{\mathbb{P}^{\lambda}[\mathcal{S}_1 > \mathcal{S}_2]}{\mathbb{E}^{\lambda}[\mathcal{S}]\mathbb{E}^{\lambda}[e^{-s\mathcal{S}}]} = \frac{2}{7}\lambda^2.$$

Let us summarize our analysis in the next remark.

Remark 4: Our analysis proposes the following analytic formula between the decision error probability and the decision delay $\mathcal{S}$:

$$\text{Prob}\,\{\text{Wrong decision}\} = \frac{2}{7}\lambda^2 \times \mathbb{E}^{\lambda}[\mathcal{S}] \times \mathbb{E}^{\lambda}[e^{-6\lambda^2\mathcal{S}}].$$

4.2.2 Pdf computation

We now come to the last part which is the computation of the pdf $q^{\lambda}(t) = \mathbb{P}^{\lambda}[\mathcal{S} \in dt]$. Since we have available the moment generating functions in (35) it suffices, as in the case $\mu = 0$, to Laplace invert it. For notational simplicity let us call $\mathcal{Q}^{\lambda}(s) = \mathbb{E}^{\lambda}[e^{s\mathcal{S}}]$ and invert the function $\mathcal{Q}^{\lambda}(s)$ using the same inverse Fourier transform technique we applied for $\mu = 0$. If in the formula (35) we replace $s = j\Omega$ then the form of the Fourier transform $\mathcal{Q}^{\lambda}(j\Omega)$ of the desired pdf is very similar to the case $\mathcal{Q}^{0}(j\Omega)$ depicted in Fig. 2. Inverting the corresponding function numerically yields the desired pdf $q^{\lambda}(t)$.

In Fig. 8 we present the result of the numerical inverse transformation for the case $\lambda = 1$, $\nu = 5$. The pdf $q^{\lambda}(t)$ is depicted in blue and for comparison we have also included the pdf $q^{0}(t)$ (black). We can see that the first pdf is concentrated on much lower values of t while the second has much thicker tails. Regarding $q^\lambda(t)$, as we can see it has a rather interesting form. The left tail seems extremely weak resembling the tails of a Gaussian, however the right tail is clearly exponential. This rather peculiar performance has been observed experimentally in behavior research (reference ????). This is a rather convincing indication that our proposed 2-CUSUM model is very efficient regarding the application of interest. As far as the exponential tail is concerned, let us again try to identify it more accurately by using the series expansion method we employed in the case of $q^0(t)$.

4.2.3 Leading exponential component

As for $\mathcal{Q}^0(s)$, we would like to apply a similar series decomposition on $\mathcal{Q}^\lambda(s)$, that is,

$$\mathcal{Q}^\lambda(s) = \sum_{k=1}^{\infty} \frac{\mathcal{B}_k}{s-\sigma_k}; \quad \text{where } \mathcal{B}_k = \lim_{s\to\sigma_k}(s-\sigma_k)\mathcal{Q}^\lambda(s). \tag{41}$$

Here $\{\sigma_k\}$ is the collection of poles for the function $\mathcal{Q}^\lambda(s)$ and $\{\mathcal{B}_k\}$ the corresponding collection of residuals. As before we assume that all the poles are simple. Inverting the previous formula yields

$$q^\lambda(t) = \sum_{k=1}^{\infty} \mathcal{B}_k e^{\sigma_k t}. \tag{42}$$

It is clear that our first step consists in finding the poles $\{\sigma_k\}$. This task turns out to be more involved than in the $\mu = 0$ case, basically because the two functions $\psi_1(0,\lambda)$ and $\psi_2(0,\lambda)$ differ significantly (while for $\mu = 0$ they are equal). By consulting (35) we realize that the poles are the roots of the equation

$$\psi_1(0,\lambda)\psi_2(0,\lambda) - 1 = 0,$$

that belong to the negative half plane (the solution $s = 0$ is excluded since for this value the numerator of $\mathcal{Q}^\lambda(s)$ is also 0 and one can verify that $s = 0$ is not a pole). If we replace the functions $\psi_i(0,\lambda)$ from (30) and (32) we end up with the equation $A(s) = 0$ where

$$A(s) = \\ \left\{\cosh\left(\sqrt{1+\frac{8s}{\lambda^2}}\frac{\nu}{2}\right) + \frac{\sinh\left(\sqrt{1+\frac{8s}{\lambda^2}}\frac{\nu}{2}\right)}{\sqrt{1+\frac{8s}{\lambda^2}}}\right\}\left\{\cosh\left(\sqrt{9+\frac{8s}{\lambda^2}}\frac{\nu}{2}\right) - 3\frac{\sinh\left(\sqrt{9+\frac{8s}{\lambda^2}}\frac{\nu}{2}\right)}{\sqrt{9+\frac{8s}{\lambda^2}}}\right\} - e^{-\nu}. \tag{43}$$

Using Hospital's rule, we can now obtain a more convenient formula for $\mathcal{B}_k$ in terms of the function $A(s)$

$$\mathcal{B}_k = -e^{-\nu} \frac{2 - \psi_1(0, \lambda) - \psi_2(0, \lambda)}{A'(s)} \Big|_{s=\sigma_k}. \tag{44}$$

For the roots of the equation $A(s) = 0$, we make the following conjecture:

Conjecture 2. *The equation $A(s) = 0$ with $A(s)$ defined in (43) has one root equal to 0 (which is not a pole) and all other roots are **real**, negative and strictly less than $-\lambda^2/8$.*

What is suggested with the conjecture is that there is no pole which lies in the interval $[-\lambda^2/8, 0)$ (as in the case $\mu = 0$) and that all poles lie in the interval $(-\infty, -\lambda^2/8)$. As long as the poles are well separated which, as we are going to see in the next part, happens when the threshold takes upon small to moderate values, the right-tail behavior is exponential and governed by the leading pole. Let us therefore compare $q^\lambda(t)$ with the leading exponential term $\mathcal{B}_1 e^{\sigma_1 t}$ and verify how well the latter describes the right-tail behavior of the pdf. In Fig. 9 we can see $q^\lambda(t)$ for $\lambda = 1$ and $\nu = 5$ at the logarithmic scale. The leading pole can be found numerically and has a value $\sigma_1 = -0.2384$ while the corresponding residue according to (44) is $\mathcal{B}_1 = 0.6613$. We can see again the perfect match between the leading exponential term and the tail of $q^\lambda(t)$.

4.2.4 Asymptotic performance

Fig. 10 depicts the typical form of the function $A(s)$. We observe that as ν increases the leading pole approaches $-\lambda^2/8$. In fact we have the following interesting asymptotic formula for this pole

$$\sigma_1 = -\frac{\lambda^2}{8} \left(1 + \frac{2\pi}{\nu} + \frac{4\pi}{\nu(\nu+2)}\right)^2.$$

From the same figure we also realize that *all* poles approach the same limit $-\lambda^2/8$ as $\nu \to \infty$. This has a very interesting consequence. Even though, for every finite value of ν, the right-tail of the pdf has an exponential profile, because all the poles tend to accumulate as $\nu \to \infty$, the tails gradually lose their initial form and *converge to a Gaussian tail.* This can be clearly seen in Fig. 11 where we present the pdf of the normalized version $\hat{\mathcal{S}} = (\mathcal{S} - \mathbb{E}^\lambda[\mathcal{S}])/\sqrt{\nu}$ of the 2-CUSUM stopping time $\mathcal{S}$ for different values of ν. This phenomenon, as we said, occurs because all poles tend to *accumulate* at $-\lambda^2/8$ suggesting that, in the limit, the expansion proposed in (41) is no longer valid. We recall that in the case $\mu = 0$ a similar accumulation takes place as well, but the leading pole remains well separated from the other poles, thus imposing an exponential behavior for $q^0(t)$, even at the limit.

Borrowing again ideas from Taylor (1975), we can demonstrate that the limiting distribution of the normalized stopping time $\hat{\mathcal{S}}$ is indeed Gaussian. This can be seen by computing the moment generating function of $\hat{\mathcal{S}} = (\mathcal{S} - \mathbb{E}^\lambda[\mathcal{S}])/\sqrt{\nu} = (\mathcal{S} - 2\nu/\lambda^2)/\sqrt{\nu} + o(1)$. Replacing s with $s/\sqrt{\nu}$, then fixing s, we can show that

$$\psi_1(0,\lambda) = e^{-\frac{2s\sqrt{\nu}}{\lambda^2}+\frac{4s^2}{\lambda^4}+o(1)}; \quad \psi_2(0,\lambda) = e^{-3\nu+O(1)}.$$

If we consider now $\mathbb{E}[e^{-\frac{s}{\sqrt{\nu}}\mathcal{S}}]$ and use (35), we can show that

$$\mathbb{E}[e^{-\frac{s}{\sqrt{\nu}}\mathcal{S}}] = \psi_1(0,\lambda)(1+o(1)) = e^{-\frac{2s\sqrt{\nu}}{\lambda^2}+\frac{4s^2}{\lambda^4}+o(1)},$$

suggesting that

$$\mathbb{E}[e^{s\hat{\mathcal{S}}}] = e^{\frac{4}{\lambda^4}s^2+o(1)}.$$

This clearly means that $\hat{\mathcal{S}}$ tends, in distribution, to a Gaussian with mean 0 and variance $4/\lambda^4$.

The limiting Gaussian behavior has of course its practical and theoretical significance, but we must keep in mind that, for small and moderate values of ν, the form of the right-tail *is clearly exponential.* Gaussian behavior is manifested only for large values of ν. In fact the exponential right-tail behavior, as we said, is very important for the application of interest since it was observed experimentally.

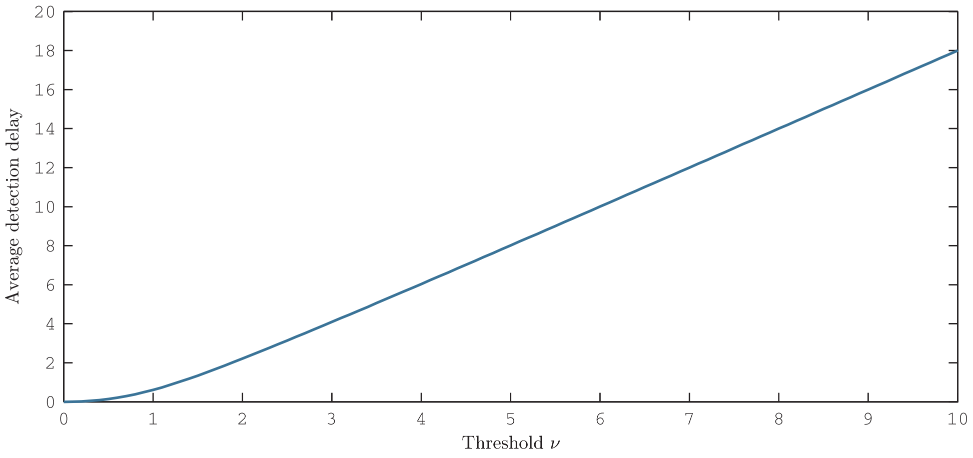

Figure 6: Average detection delay as a function of the threshold ν.

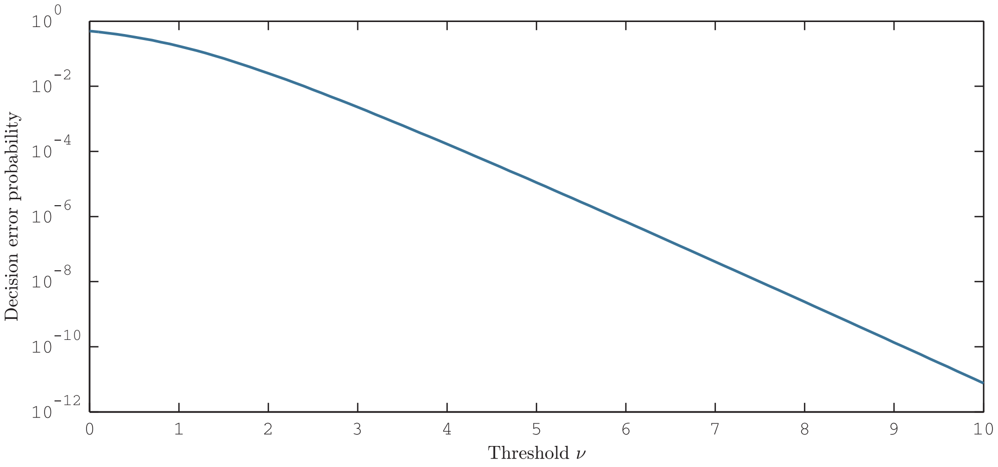

Figure 7: Decision error probability as a function of the threshold ν.

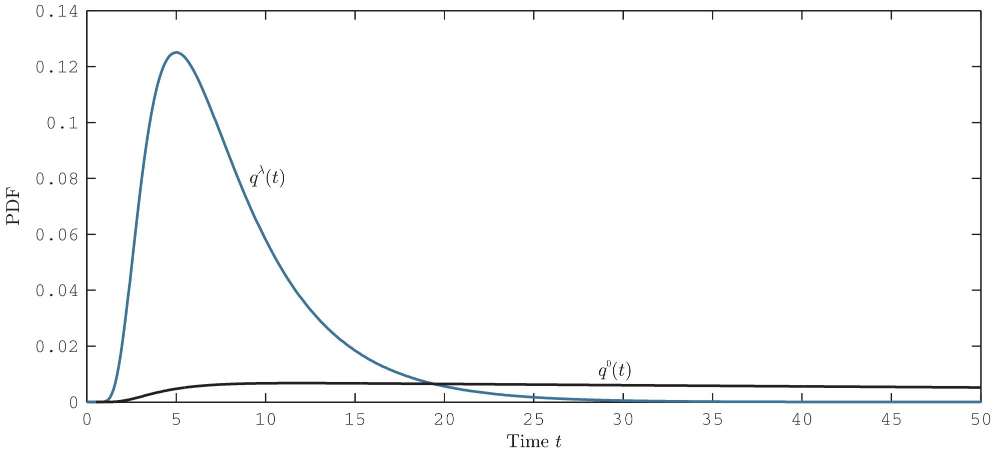

Figure 8: Pdf $q^\lambda(t)$ (blue) and $q^0(t)$ (black) for $\lambda = 1$ and $\nu = 5$.

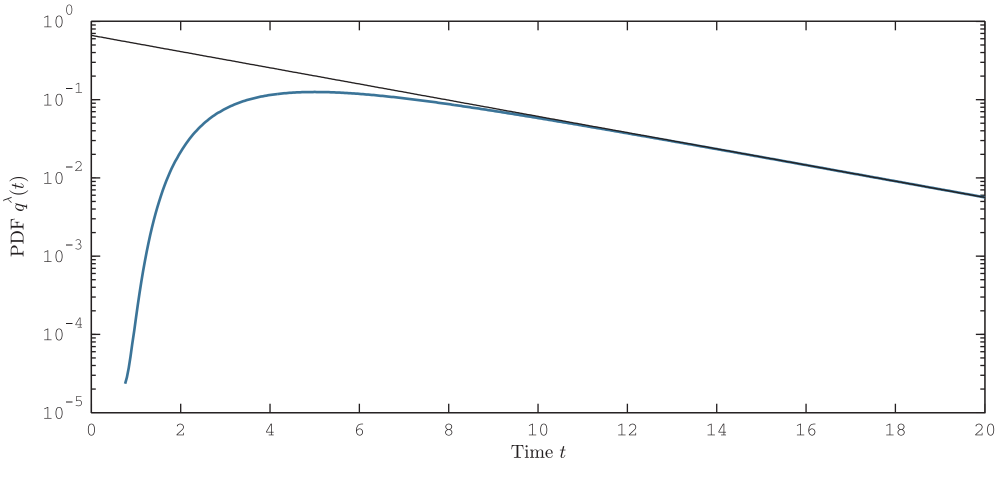

Figure 9: Pdf $q^\lambda(t)$ and corresponding leading exponential component.

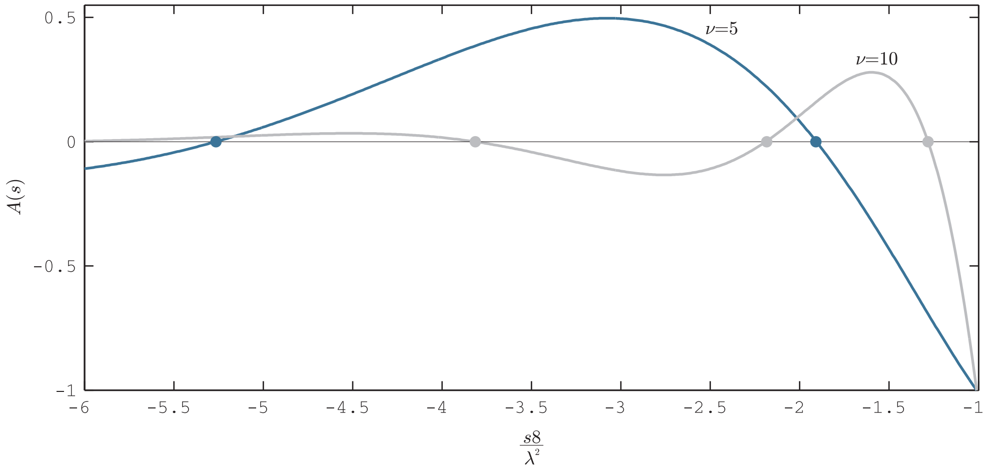

Figure 10: Form of the function $A(s)$ and corresponding poles for different values of the threshold ν.

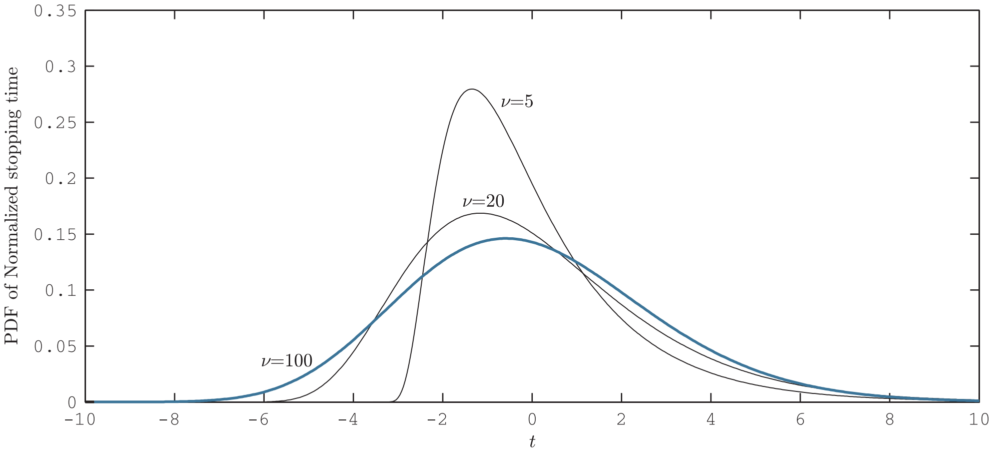

Figure 11: Pdf of the normalized 2-CUSUM stopping time $\hat{\mathcal{S}} = (\mathcal{S} - \mathbb{E}^{\lambda}[\mathcal{S}])/\sqrt{\nu}$, for different values of the threshold ν.